Parallelized Stochastic Gradient Descent

Martin A. Zinkevich
Yahoo! Labs
Sunnyvale, CA 94089
maz@yahoo-inc.com

Markus Weimer
Yahoo! Labs
Sunnyvale, CA 94089
weimer@yahoo-inc.com

Alex Smola
Yahoo! Labs
Sunnyvale, CA 94089
smola@yahoo-inc.com

Lihong Li
Yahoo! Labs
Sunnyvale, CA 94089
lihong@yahoo-inc.com

Abstract

With the increase in available data parallel machine learning has become an increasingly pressing problem. In this paper we present the first parallel stochastic gradient descent algorithm including a detailed analysis and experimental evidence. Unlike prior work on parallel optimization algorithms [5, 7] our variant comes with parallel acceleration guarantees and it poses no overly tight latency constraints, which might only be available in the multicore setting. Our analysis introduces a novel proof technique — contractive mappings to quantify the speed of convergence of parameter distributions to their asymptotic limits. As a side effect this answers the question of how quickly stochastic gradient descent algorithms reach the asymptotically normal regime [1, 8].

1 Introduction

Over the past decade the amount of available data has increased steadily. By now some industrial scale datasets are approaching Petabytes. Given that the bandwidth of storage and network per computer has not been able to keep up with the increase in data, the need to design data analysis algorithms which are able to perform most steps in a distributed fashion without tight constraints on communication has become ever more pressing. A simple example illustrates the dilemma. At current disk bandwidth and capacity (2TB at 100MB/s throughput) it takes at least 6 hours to read the content of a single harddisk. For a decade, the move from batch to online learning algorithms was able to deal with increasing data set sizes, since it reduced the runtime behavior of inference algorithms from cubic or quadratic to linear in the sample size. However, whenever we have more than a single disk of data, it becomes computationally infeasible to process all data by stochastic gradient descent which is an inherently sequential algorithm, at least if we want the result within a matter of hours rather than days.

Three recent papers attempted to break this parallelization barrier, each of them with mixed success. [5] show that parallelization is easily possible for the *multicore* setting where we have a tight coupling of the processing units, thus ensuring extremely low latency between the processors. In particular, for non-adversarial settings it is possible to obtain algorithms which scale perfectly in the number of processors, both in the case of bounded gradients and in the strongly convex case. Unfortunately, these algorithms are not applicable to a MapReduce setting since the latter is fraught with considerable latency and bandwidth constraints between the computers.

A more MapReduce friendly set of algorithms was proposed by [3, 9]. In a nutshell, they rely on distributed computation of gradients locally on each computer which holds parts of the data and subsequent aggregation of gradients to perform a global update step. This algorithm scales linearly

in the amount of data and log-linearly in the number of computers. That said, the overall cost in terms of computation and network is very high: it requires many passes through the dataset for convergence. Moreover, it requires many synchronization sweeps (i.e. MapReduce iterations). In other words, this algorithm is computationally very wasteful when compared to online algorithms.

[7] attempted to deal with this issue by a rather ingenious strategy: solve the sub-problems exactly on each processor and in the end average these solutions to obtain a joint solution. The key advantage of this strategy is that only a single MapReduce pass is required, thus dramatically reducing the amount of communication. Unfortunately their proposed algorithm has a number of drawbacks: the theoretical guarantees they are able to obtain imply a significant *variance* reduction relative to the single processor solution [7, Theorem 3, equation 13] but *no bias reduction whatsoever* [7, Theorem 2, equation 9] relative to a single processor approach. Furthermore, their approach requires a relatively expensive algorithm (a full batch solver) to run on each processor. A further drawback of the analysis in [7] is that the convergence guarantees are very much dependent on the degree of strong convexity as endowed by regularization. However, since regularization tends to decrease with increasing sample size the guarantees become increasingly loose in practice as we see more data.

We attempt to combine the benefits of a single-average strategy as proposed by [7] with asymptotic analysis [8] of online learning. Our proposed algorithm is strikingly simple: denote by $c^i(w)$ a loss function indexed by i and with parameter w. Then each processor carries out stochastic gradient descent on the set of $c^i(w)$ with a fixed learning rate η for T steps as described in Algorithm 1.

Algorithm 1 $\mathrm{SGD}(\{c^1, \ldots, c^m\}, T, \eta, w_0)$

for $t = 1$ **to** T **do**
  Draw $j \in \{1 \ldots m\}$ uniformly at random.
  $w_t \leftarrow w_{t-1} - \eta \partial_w c^j(w_{t-1})$.
end for
return w_T.

On top of the SGD routine which is carried out on each computer we have a master-routine which aggregates the solution in the same fashion as [7].

Algorithm 2 $\mathrm{ParallelSGD}(\{c^1, \ldots c^m\}, T, \eta, w_0, k)$

for all $i \in \{1, \ldots k\}$ **parallel do**
  $v_i = \mathrm{SGD}(\{c^1, \ldots c^m\}, T, \eta, w_0)$ on client
end for
Aggregate from all computers $v = \frac{1}{k} \sum_{i=1}^{k} v_i$ and **return** v

The key *algorithmic* difference to [7] is that the batch solver of the inner loop is replaced by a stochastic gradient descent algorithm which digests *not* a fixed fraction of data but rather a random fixed subset of data. This means that if we process T instances per machine, each processor ends up seeing $\frac{T}{m}$ of the data which is likely to exceed $\frac{1}{k}$.

Algorithm	Latency tolerance	MapReduce	Network IO	Scalability
Distributed subgradient [3, 9]	moderate	**yes**	high	**linear**
Distributed convex solver [7]	**high**	**yes**	**low**	unclear
Multicore stochastic gradient [5]	low	no	n.a.	**linear**
This paper	**high**	**yes**	**low**	**linear**

A direct implementation of the algorithms above would place every example on every machine: however, if T is much less than m, then it is only necessary for a machine to have access to the data it actually touches. Large scale learning, as defined in [2], is when an algorithm is bounded by the time available instead of by the amount of data available. Practically speaking, that means that one can consider the actual data in the real dataset to be a subset of a virtually infinite set, and drawing with replacement (as the theory here implies) and drawing without replacement on the

Algorithm 3 SimuParallelSGD(Examples $\{c^1, \dots c^m\}$, Learning Rate η, Machines k)

Define $T = \lfloor m/k \rfloor$
Randomly partition the examples, giving T examples to each machine.
for all $i \in \{1, \dots k\}$ **parallel do**
  Randomly shuffle the data on machine i.
  Initialize $w_{i,0} = 0$.
  for all $t \in \{1, \dots T\}$: **do**
    Get the tth example on the ith machine (this machine), $c^{i,t}$
    $w_{i,t} \leftarrow w_{i,t-1} - \eta \partial_w c^i(w_{i,t-1})$
  end for
end for
Aggregate from all computers $v = \frac{1}{k} \sum_{i=1}^{k} w_{i,t}$ and **return** v.

infinite data set can both be simulated by shuffling the real data and accessing it sequentially. The initial distribution and shuffling can be a part of how the data is saved. SimuParallelSGD fits very well with the large scale learning paradigm as well as the MapReduce framework. Our paper applies an anytime algorithm via stochastic gradient descent. The algorithm requires no communication between machines until the end. This is perfectly suited to MapReduce settings. Asymptotically, the error approaches zero. The amount of time required is independent of the number of examples, only depending upon the regularization parameter and the desired error at the end.

2 Formalism

In stark contrast to the simplicity of Algorithm 2, its convergence analysis is highly technical. Hence we limit ourselves to presenting the main results in this extended abstract. Detailed proofs are given in the appendix. Before delving into details we briefly outline the proof strategy:

- When performing stochastic gradient descent with fixed (and sufficiently small) learning rate η the distribution of the parameter vector is asymptotically normal [1, 8]. Since all computers are drawing from the same data distribution they all converge to the same limit.
- Averaging between the parameter vectors of k computers reduces variance by $O(k^{-\frac{1}{2}})$ similar to the result of [7]. However, it does *not* reduce bias (this is where [7] falls short).
- To show that the bias due to joint initialization decreases we need to show that the *distribution* of parameters per machine converges sufficiently quickly to the limit distribution.
- Finally, we also need to show that the mean of the limit distribution for fixed learning rate is sufficiently close to the risk minimizer. That is, we need to take finite-size learning rate effects into account relative to the asymptotically normal regime.

2.1 Loss and Contractions

In this paper we consider estimation with convex loss functions $c^i : \ell_2 \to [0, \infty)$. While our analysis extends to other Hilbert Spaces such as RKHSs we limit ourselves to this class of functions for convenience. For instance, in the case of regularized risk minimization we have

$$c^i(w) = \frac{\lambda}{2}\|w\|^2 + L(x^i, y^i, w \cdot x^i) \tag{1}$$

where L is a convex function in $w \cdot x^i$, such as $\frac{1}{2}(y^i - w \cdot x^i)^2$ for regression or $\log[1 + \exp(-y^i w \cdot x^i)]$ for binary classification. The goal is to find an approximate minimizer of the overall risk

$$c(w) = \frac{1}{m} \sum_{i=1}^{m} c^i(w). \tag{2}$$

To deal with *stochastic* gradient descent we need tools for quantifying distributions over w.

Lipschitz continuity: A function $f : \mathcal{X} \to \mathbf{R}$ is Lipschitz continuous with constant L with respect to a distance d if $|f(x) - f(y)| \leq L d(x, y)$ for all $x, y \in \mathcal{X}$.

Hölder continuity: A function f is Hölder continous with constant L and exponent α if $|f(x) - f(y)| \leq Ld^{\alpha}(x, y)$ for all $x, y \in \mathcal{X}$.

Lipschitz seminorm: [10] introduce a seminorm. With minor modification we use

$$\|f\|_{\text{Lip}} := \inf \{l \text{ where } |f(x) - f(y)| \leq ld(x, y) \text{ for all } x, y \in \mathcal{X}\}. \tag{3}$$

That is, $\|f\|_{\text{Lip}}$ is the smallest constant for which Lipschitz continuity holds.

Hölder seminorm: Extending the Lipschitz norm for $\alpha \geq 1$:

$$\|f\|_{\text{Lip}_\alpha} := \inf \{l \text{ where } |f(x) - f(y)| \leq ld^{\alpha}(x, y) \text{ for all } x, y \in \mathcal{X}\}. \tag{4}$$

Contraction: For a metric space (M, d), $f : M \to M$ is a contraction mapping if $\|f\|_{\text{Lip}} < 1$.

In the following we assume that $\|L(x, y, y')\|_{\text{Lip}} \leq G$ as a function of y' for all occurring data $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and for all values of w within a suitably chosen (often compact) domain.

Theorem 1 (Banach's Fixed Point Theorem) *If (M, d) is a non-empty complete metric space, then any contraction mapping f on (M, d) has a unique fixed point $x^* = f(x^*)$.*

Corollary 2 *The sequence $x_t = f(x_{t-1})$ converges linearly with $d(x^*, x_t) \leq \|f\|^t_{\text{Lip}} d(x_0, x^*)$.*

Our strategy is to show that the stochastic gradient descent mapping

$$w \leftarrow \phi^i(w) := w - \eta \nabla c^i(w) \tag{5}$$

is a contraction, where i is selected uniformly at random from $\{1, \dots m\}$. This would allow us to demonstrate exponentially fast convergence. Note that since the algorithm selects i at random, different runs with the same initial settings can produce different results. A key tool is the following:

Lemma 3 *Let $c^* \geq \left\|\partial_{\hat{y}} L(x^i, y^i, \hat{y})\right\|_{\text{Lip}}$ be a Lipschitz bound on the loss gradient. Then if $\eta \leq (\left\|x^i\right\|^2 c^* + \lambda)^{-1}$ the update rule (5) is a contraction mapping in ℓ_2 with Lipschitz constant $1 - \eta\lambda$.*

We prove this in Appendix B. If we choose η "low enough", gradient descent uniformly becomes a contraction. We define

$$\eta^* := \min_i \left(\left\|x^i\right\|^2 c^* + \lambda\right)^{-1}. \tag{6}$$

2.2 Contraction for Distributions

For fixed learning rate η stochastic gradient descent is a Markov process with state vector w. While there is considerable research regarding the asymptotic properties of this process [1, 8], not much is known regarding the number of iterations required until the asymptotic regime is assumed. We now address the latter by extending the notion of contractions from mappings of points to mappings of distributions. For this we introduce the Monge-Kantorovich-Wasserstein earth mover's distance.

Definition 4 (Wasserstein metric) *For a Radon space (M, d) let $P(M, d)$ be the set of all distributions over the space. The Wasserstein distance between two distributions $X, Y \in P(M, d)$ is*

$$W_z(X, Y) = \left[\inf_{\gamma \in \Gamma(X,Y)} \int_{x,y} d^z(x, y) d\gamma(x, y)\right]^{\frac{1}{z}} \tag{7}$$

where $\Gamma(X, Y)$ is the set of probability distributions on $(M, d) \times (M, d)$ with marginals X and Y.

This metric has two very important properties: it is complete and a contraction in (M, d) induces a contraction in $(P(M, d), W_z)$. Given a mapping $\phi : M \to M$, we can construct $\mathbf{p} : P(M, d) \to P(M, d)$ by applying ϕ pointwise to M. Let $X \in P(M, d)$ and let $X' := \mathbf{p}(X)$. Denote for any measurable event E its pre-image by $\phi^{-1}(E)$. Then we have that $X'(E) = X(\phi^{-1}(E))$.

Lemma 5 *Given a metric space (M, d) and a contraction mapping ϕ on (M, d) with constant c, $\mathbf{p}$ is a contraction mapping on $(P(M, d), W_z)$ with constant c.*

This is proven in Appendix C. This shows that any single mapping is a contraction. However, since we draw c^i at random we need to show that a mixture of such mappings is a contraction, too. Here the fact that we operate on distributions comes handy since the mixture of mappings on distribution is a mapping on distributions.

Lemma 6 *Given a Radon space (M, d), if $\mathbf{p}_1 \dots \mathbf{p}_k$ are contraction mappings with constants $c_1 \dots c_k$ with respect to W_z, and $\sum_i a_i = 1$ where $a_i \geq 0$, then $\mathbf{p} = \sum_{i=1}^k a_i \mathbf{p}_i$ is a contraction mapping with a constant of no more than $[\sum_i a_i (c_i)^z]^{\frac{1}{z}}$.*

Corollary 7 *If for all i, $c_i \leq c$, then $\mathbf{p}$ is a contraction mapping with a constant of no more than c.*

This is proven in Appendix C. We apply this to SGD as follows: Define $\mathbf{p}^* = \frac{1}{m} \sum_{i=1}^m \mathbf{p}^i$ to be the stochastic operation in one step. Denote by D_η^0 the initial parameter distribution from which w_0 is drawn and by D_η^t the parameter distribution after t steps, which is obtained via $D_\eta^t = \mathbf{p}^*(D_\eta^{t-1})$. Then the following holds:

Theorem 8 *For any $z \in \mathbf{N}$, if $\eta \leq \eta^*$, then $\mathbf{p}^*$ is a contraction mapping on (M, W_z) with contraction rate $(1 - \eta\lambda)$. Moreover, there exists a unique fixed point D_η^* such that $\mathbf{p}^*(D_\eta^*) = D_\eta^*$. Finally, if $w_0 = 0$ with probability 1, then $W_z(D_\eta^0, D_\eta^*) = \frac{G}{\lambda}$, and $W_z(D_\eta^T, D_\eta^*) \leq \frac{G}{\lambda}(1 - \eta\lambda)^T$.*

This is proven in Appendix F. The contraction rate $(1 - \eta\lambda)$ can be proven by applying Lemma 3, Lemma 5, and Corollary 6. As we show later, $w_t \leq G/\lambda$ with probability 1, so $\Pr_{w \in D_\eta^*}[d(0, w) \leq G/\lambda] = 1$, and since $w_0 = 0$, this implies $W_z(D_\eta^0, D_\eta^*) = G/\lambda$. From this, Corollary 2 establishes $W_z(D_\eta^T, D_\eta^*) \leq \frac{G}{\lambda}(1 - \eta\lambda)^T$.

This means that for a suitable choice of η we achieve exponentially fast convergence in T to some stationary distribution D_η^*. Note that this distribution need *not* be centered at the risk minimizer of $c(w)$. What the result does, though, is establish a guarantee that each computer carrying out Algorithm 1 will converge rapidly to the same distribution over w, which will allow us to obtain good bounds if we can bound the 'bias' and 'variance' of D_η^*.

2.3 Guarantees for the Stationary Distribution

At this point, we know there exists a stationary distribution, and our algorithms are converging to that distribution exponentially fast. However, unlike in traditional gradient descent, the stationary distribution is not necessarily just the optimal point. In particular, the harder parts of understanding this algorithm involve understanding the properties of the stationary distribution. First, we show that the mean of the stationary distribution has low error. Therefore, if we ran for a really long time and averaged over many samples, the error would be low.

Theorem 9 $c(\mathbf{E}_{w \in D_\eta^*}[w]) - \min_{w \in \mathbf{R}^n} c(w) \leq 2\eta G^2$.

Proven in Appendix G using techniques from regret minimization. Secondly, we show that the squared distance from the optimal point, and therefore the variance, is low.

Theorem 10 *The average squared distance of D_η^* from the optimal point is bounded by:*

$$\mathbf{E}_{w \in D_\eta^*}[(w - w^*)^2] \leq \frac{4\eta G^2}{(2 - \eta\lambda)\lambda}.$$

In other words, the squared distance is bounded by $O(\eta G^2/\lambda)$.

Proven in Appendix I using techniques from reinforcement learning. In what follows, if $x \in M$, $Y \in P(M, d)$, we define $W_z(x, Y)$ to be the W_z distance between Y and a distribution with a probability of 1 at x. Throughout the appendix, we develop tools to show that the distribution over the output vector of the algorithm is "near" $\mu_{D_\eta^*}$, the mean of the stationary distribution. In particular, if $D_\eta^{T,k}$ is the distribution over the final vector of $\mathrm{ParallelSGD}$ after T iterations on each of k machines with a learning rate η, then $W_2(\mu_{D_\eta^*}, D_\eta^{T,k}) = \sqrt{\mathbf{E}_{x \in D_\eta^{T,k}}[(x - \mu_{D_\eta^*})^2]}$ becomes small. Then, we need to connect the error of the mean of the stationary distribution to a distribution that is near to this mean.

Theorem 11 *Given a cost function c such that $\|c\|_L$ and $\|\nabla c\|_L$ are bounded, a distribution D such that σ_D and is bounded, then, for any v:*

$$\mathbf{E}_{w \in D}[c(w)] - \min_w c(w) \leq (W_2(v, D)) \sqrt{2 \|\nabla c\|_L (c(v) - \min_w c(w))} + \frac{\|\nabla c\|_L}{2} (W_2(v, D))^2 + (c(v) - \min_w c(w)) \tag{8}$$

This is proven in Appendix K. The proof is related to the Kantorovich-Rubinstein theorem, and bounds on the Lipschitz of c near v based on $c(v) - \min_w c(w)$. At this point, we are ready to get the *main theorem:*

Theorem 12 *If $\eta \leq \eta^*$ and $T = \frac{\ln k - (\ln \eta + \ln \lambda)}{2\eta\lambda}$:*

$$\mathbf{E}_{w \in D_\eta^{T,k}}[c(w)] - \min_w c(w) \leq \frac{8\eta G^2}{\sqrt{k\lambda}} \sqrt{\|\nabla c\|_L} + \frac{8\eta G^2 \|\nabla c\|_L}{k\lambda} + (2\eta G^2). \tag{9}$$

This is proven in Appendix K.

2.4 Discussion of the Bound

The guarantee obtained in (9) appears rather unusual insofar as it does not have an explicit dependency on the sample size. This is to be expected since we obtained a bound in terms of risk minimization of the given corpus rather than a learning bound. Instead the runtime required depends only on the accuracy of the solution itself.

In comparison to [2], we look at the number of iterations to reach ρ for SGD in Table 2. Ignoring the effect of the dimensions (such as ν and d), setting these parameters to 1, and assuming that the conditioning number $\kappa = \frac{1}{\lambda}$, and $\rho = \eta$. In terms of our bound, we assume $G = 1$ and $\|\nabla c\|_L = 1$. In order to make our error order η, we must set $k = \frac{1}{\lambda}$. So, the Bottou paper claims a bound of $\frac{\nu\kappa^2}{\rho}$ iterations, which we interpret as $\frac{1}{\eta\lambda^2}$. Modulo logarithmic factors, we require $\frac{1}{\lambda}$ machines to run $\frac{1}{\eta\lambda}$ time, which is the same order of computation, but a dramatic speedup of a factor of $\frac{1}{\lambda}$ in wall clock time.

Another important aspect of the algorithm is that it can be arbitrarily precise. By halving η and roughly doubling T, you can halve the error. Also, the bound captures how much paralllelization can help. If $k > \frac{\|\nabla c\|_L}{\lambda}$, then the last term ηG^2 will start to dominate.

3 Experiments

Data: We performed experiments on a proprietary dataset drawn from a major email system with labels $y \in \pm 1$ and binary, sparse features. The dataset contains $3,189,235$ time-stamped instances out of which the last $68,1015$ instances are used to form the test set, leaving $2,508,220$ training points. We used hashing to compress the features into a 2^{18} dimensional space. In total, the dataset contained $785,751,531$ features after hashing, which means that each instance has about 313 features on average. Thus, the average sparsity of each data point is 0.0012. All instance have been normalized to unit length for the experiments.

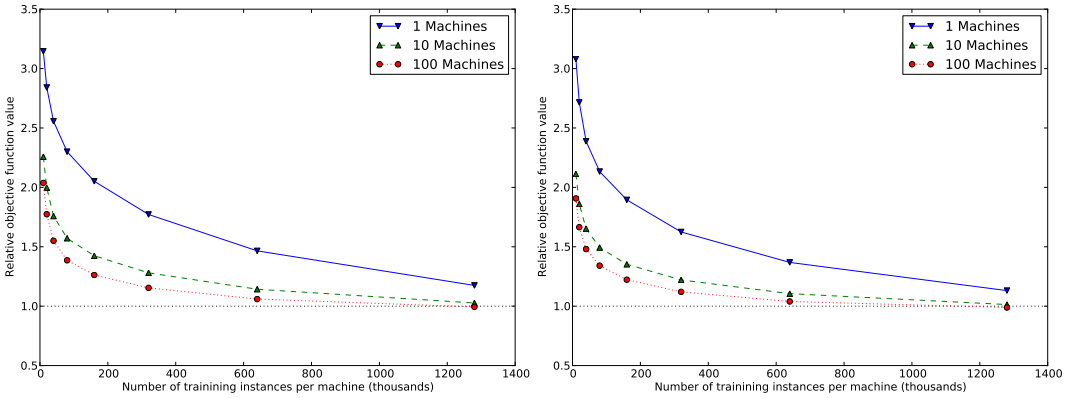

Figure 1: Relative training error with $\lambda = 1e^{-3}$: Huber loss (left) and squared error (right)

Approach: In order to evaluate the parallelization ability of the proposed algorithm, we followed the following procedure: For each configuration (see below), we trained up to 100 models, each on an independent, random permutation of the full training data. During training, the model is stored on disk after $k = 10,000 * 2^i$ updates. We then averaged the models obtained for each i and evaluated the resulting model. That way, we obtained the performance for the algorithm after each machine has seen k samples. This approach is geared towards the estimation of the parallelization ability of our optimization algorithm and its application to machine learning equally. This is in contrast to the evaluation approach taken in [7] which focussed solely on the machine learning aspect without studying the performance of the optimization approach.

Evaluation measures: We report both the normalized root mean squared error (RMSE) on the test set and the normalized value of the objective function during training. We normalize the RMSE such that 1.0 is the RMSE obtained by training a model in one single, sequential pass over the data. The objective function values are normalized in much the same way such that the objective function value of a single, full sequential pass over the data reaches the value 1.0.

Configurations: We studied both the Huber and the squared error loss. While the latter does not satisfy all the assumptions of our proofs (its gradient is unbounded), it is included due to its popularity. We choose to evaluate using two different regularization constants, $\lambda = 1e^{-3}$ and $\lambda = 1e^{-6}$ in order to estimate the performance characteristics both on smooth, "easy" problems ($1e^{-3}$) and on high-variance, "hard" problems ($1e^{-6}$). In all experiments, we fixed the learning rate to $\eta = 1e^{-3}$.

3.1 Results and Discussion

Optimization: Figure 1 shows the relative objective function values for training using $1, 10$ and 100 machines with $\lambda = 1e^{-3}$. In terms of *wall clock time*, the models obtained on 100 machines clearly outperform the ones obtained on 10 machines, which in turn outperform the model trained on a single machine. There is no significant difference in behavior between the squared error and the Huber loss in these experiments, despite the fact that the squared error is effectively unbounded. Thus, the parallelization works in the sense that many machines obtain a better objective function value after each machine has seen k instances. Additionally, the results also show that data-local parallelized training is feasible and beneficial with the proposed algorithm in practice. Note that the parallel training needs slightly more *machine time* to obtain the same objective function value, which is to be expected. Also unsurprising, yet noteworthy, is the trade-off between the number of machines and the quality of the solution: The solution obtained by 10 machines is much more of an improvement over using one machine than using 100 machines is over 10.

Predictive Performance: Figure 2 shows the relative test RMSE for $1, 10$ and 100 machines with $\lambda = 1e^{-3}$. As expected, the results are very similar to the objective function comparison: The parallel training decreases *wall clock time* at the price of slightly higher *machine time*. Again, the gain in performance between 1 and 10 machines is much higher than the one between 10 and 100.

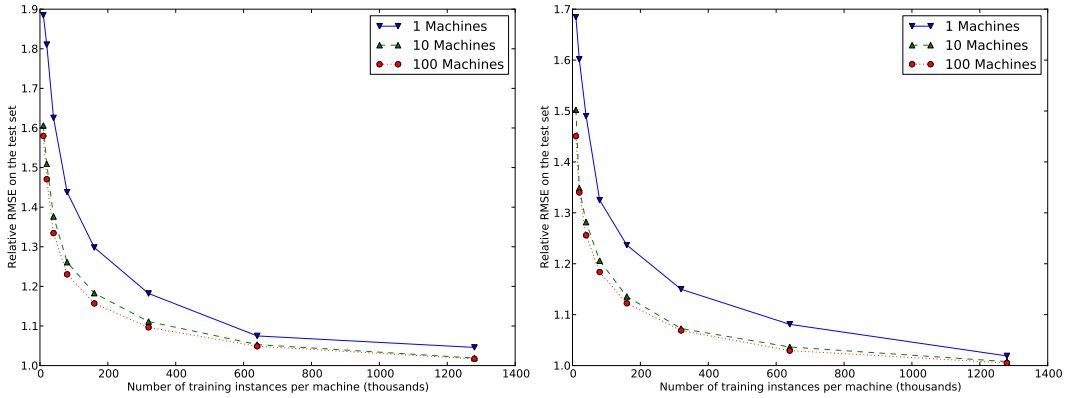

Figure 2: Relative Test-RMSE with $\lambda = 1e^{-3}$: Huber loss (left) and squared error (right)

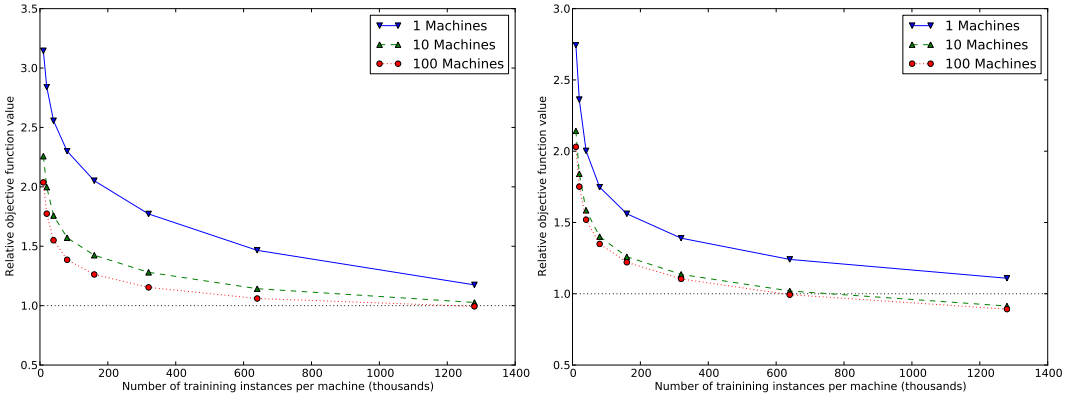

Figure 3: Relative train-error using Huber loss: $\lambda = 1e^{-3}$ (left), $\lambda = 1e^{-6}$ (right)

Performance using different λ: The last experiment is conducted to study the effect of the regularization constant λ on the parallelization ability: Figure 3 shows the objective function plot using the Huber loss and $\lambda = 1e^{-3}$ and $\lambda = 1e^{-6}$. The lower regularization constant leads to more variance in the problem which in turn should increase the benefit of the averaging algorithm. The plots exhibit exactly this characteristic: For $\lambda = 1e^{-6}$, the loss for 10 and 100 machines not only drops faster, but the final solution for both beats the solution found by a single pass, adding further empirical evidence for the behaviour predicted by our theory.

4 Conclusion

In this paper, we propose a novel *data-parallel* stochastic gradient descent algorithm that enjoys a number of key properties that make it highly suitable for parallel, large-scale machine learning: It imposes very little I/O overhead: Training data is accessed locally and only the model is communicated at the very end. This also means that the algorithm is indifferent to I/O latency. These aspects make the algorithm an ideal candidate for a MapReduce implementation. Thereby, it inherits the latter's superb data locality and fault tolerance properties. Our analysis of the algorithm's performance is based on a novel technique that uses contraction theory to quantify finite-sample convergence rate of stochastic gradient descent. We show worst-case bounds that are comparable to stochastic gradient descent in terms of wall clock time, and vastly faster in terms of overall time. Lastly, our experiments on a large-scale real world dataset show that the parallelization reduces the wall-clock time needed to obtain a set solution quality. Unsurprisingly, we also see diminishing marginal utility of adding more machines. Finally, solving problems with more variance (smaller regularization constant) benefits more from the parallelization.

References

[1] Shun-ichi Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, 16:299–307, 1967.

[2] L. Bottou and O. Bosquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, 2008.

[3] C.T. Chu, S.K. Kim, Y. A. Lin, Y. Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, 2007.

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Computational Learning Theory*, 2010.

[5] J. Langford, A.J. Smola, and M. Zinkevich. Slow learners are fast. In *Neural Information Processing Systems*, 2009.

[6] J. Langford, A.J. Smola, and M. Zinkevich. Slow learners are fast. arXiv:0911.0491, 2009.

[7] G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker. Efficient large-scale distributed training of conditional maximum entropy models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1231–1239. 2009.

[8] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion — determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.

[9] Choon Hui Teo, S. V. N. Vishwanthan, Alex J. Smola, and Quoc V. Le. Bundle methods for regularized risk minimization. *J. Mach. Learn. Res.*, 11:311–365, January 2010.

[10] U. von Luxburg and O. Bousquet. Distance-based classification with lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

[11] M. Zinkevich. Online convex programming and generalised infinitesimal gradient ascent. In *Proc. Intl. Conf. Machine Learning*, pages 928–936, 2003.

A Contraction Proof for Strongly Convex Functions

Lemma 13 *(Lemma 7, [6]) Assume that f is convex and moreover that $\nabla f(x)$ is Lipschitz continuous with constant H. Finally, denote by x^* the minimizer of f. In this case*

$$\|\nabla f(x)\|^2 \leq 2H[f(x) - f(x^*)]. \tag{10}$$

c is **λ-strongly convex** if for all $x, y \in M$:

$$\frac{\lambda}{2}(y-x)^2 + \nabla c(x) \cdot (y-x) + c(x) \leq c(y) \tag{11}$$

Lemma 14 *If c is λ-strongly convex, x^* is the minimizer of c, then $f(x) = c(x) - \frac{\lambda}{2}(x-x^*)^2$ is convex and x^* minimizes f.*

Proof Note that for $x, y \in M$:

$$\frac{\lambda}{2}(y-x)^2 + \nabla c(x) \cdot (y-x) + c(x) \leq c(y) \tag{12}$$

$$\nabla f(x) = \nabla c(x) - \lambda(x - x^*) \tag{13}$$

We can write ∇c and c as functions of f:

$$\nabla c(x) = \nabla f(x) + \lambda(x - x^*) \tag{14}$$

$$c(x) = f(x) + \frac{\lambda}{2}(x-x^*)^2 \tag{15}$$

Plugging f and ∇f into Equation 12 yields:

$$\frac{\lambda}{2}(y-x)^2 + \nabla f(x) \cdot (y-x) + \lambda(x-x^*) \cdot (y-x) + f(x) + \frac{\lambda}{2}(x-x^*)^2 \leq f(y) + \frac{\lambda}{2}(y-x^*)^2 \tag{16}$$

$$-\lambda y \cdot x + \nabla f(x) \cdot (y-x) + \lambda x \cdot y - \lambda x^* \cdot y + \lambda x \cdot x^* + f(x) - \lambda x \cdot x^* \leq f(y) - \lambda y \cdot x^* \tag{17}$$

$$\nabla f(x) \cdot (y-x) + f(x) \leq f(y) \tag{18}$$

Thus, f is convex. Moreover, since $\nabla f(x^*) = \nabla c(x^*) - \lambda(x^* - x^*) = \nabla c(x^*) = 0$, then x^* is optimal for f as well as c. ■

Lemma 15 *If c is λ-strongly convex, x^* is the minimizer of c, ∇c is Lipschitz continuous $f(x) = c(x) - \frac{\lambda}{2}(x-x^*)^2$, $\eta < \left(\lambda + \|\nabla f\|_{\text{Lip}}\right)^{-1}$, and $\eta < 1$, then for all $x \in M$:*

$$d(x - \eta \nabla c(x), x^*) \leq (1 - \eta\lambda) d(x, x^*) \tag{19}$$

Proof

To keep things terse, define $H := \|\nabla c\|_{\text{Lip}}$.

First observe that $\lambda + \|\nabla f\|_{\text{Lip}} \geq \|\nabla c\|_{\text{Lip}}$, so $\eta < H^{-1}$.

Without loss of generality, assume $x^* = 0$. By the definition of Lipschitz continuous, $\|\nabla c(x) - \nabla c(x^*)\| \leq H \|x - x^*\|$ and therefore $\|\nabla c(x)\| \leq H \|x\|$. Therefore, $\nabla c(x) \cdot x \leq H \|x\|^2$. In other words:

$$(x - \eta \nabla c(x)) \cdot x = x \cdot x - \eta \nabla c(x) \cdot x \tag{20}$$

$$(x - \eta \nabla c(x)) \cdot x \geq \|x\|^2 (1 - \eta H) \tag{21}$$

Therefore, at least in the direction of x, if $\eta < H^{-1}$, then $(x - \eta \nabla c(x)) \cdot x \geq 0$. Define $H' = \|\nabla f\|_{\text{Lip}}$. Since f is convex and x^* is optimal:

$$\nabla f(x) \cdot (0 - x) + f(x) \leq f(x^*) \tag{22}$$
$$f(x) - f(x^*) \leq \nabla f(x) \cdot x \tag{23}$$
$$\tag{24}$$

By Lemma 13:

$$\frac{\|\nabla f(x)\|^2}{2H'} \leq \nabla f(x) \cdot x \tag{25}$$

We break down $\nabla f(x)$ into $g_\|$ and $g_\perp$, such that $g_\| = \frac{\nabla f(x) \cdot x}{\|x\|^2} x$, and $g_\perp = x - g_\|$. Therefore, $g_\perp \cdot g_\| = 0$, and $\|\nabla f(x)\|^2 = \|g_\|\|^2 + \|g_\perp\|^2$, and $\nabla c(x) \cdot x = (\lambda x + g_\|) \cdot x$. Thus, since we know $(x - \eta \nabla c(x)) \cdot x$ is positive, we can write:

$$\|x - \eta \nabla c(x)\|^2 = \|x - \eta \lambda x - \eta g_\|\|^2 + \|\eta g_\perp\|^2 \tag{26}$$

Thus, looking at $\|(1 - \eta\lambda)x - \eta g_\|\|^2$:

$$\|(1 - \eta\lambda)x - \eta g_\|\|^2 = ((1 - \eta\lambda)x - \eta g_\|) \cdot ((1 - \eta\lambda)x - \eta g_\|) \tag{27}$$
$$\|(1 - \eta\lambda)x - \eta g_\|\|^2 = (1 - \eta\lambda)^2 \|x\|^2 - 2(1 - \eta\lambda)\eta g_\| \cdot x + \eta^2 \|g_\|\|^2 \tag{28}$$
$$\|(1 - \eta\lambda)x - \eta g_\|\|^2 \leq (1 - \eta\lambda)^2 \|x\|^2 - 2(1 - \eta\lambda)\frac{\|\nabla f(x)\|^2}{2H'} + \eta^2 \|g_\|\|^2 \tag{29}$$
$$\|(1 - \eta\lambda)x - \eta g_\|\|^2 \leq (1 - \eta\lambda)^2 \|x\|^2 - 2(1 - \eta\lambda)\frac{\|g_\|\|^2 + \|g_\perp\|^2}{2H'} + \eta^2 \|g_\|\|^2 \tag{30}$$

$$\|x - \eta \nabla c\|^2 \leq (1 - \eta\lambda)^2 \|x\|^2 - 2(1 - \eta\lambda)\frac{\|g_\|\|^2 + \|g_\perp\|^2}{2H'} + \eta^2 \|g_\|\|^2 + \|\eta g_\perp\|^2 \tag{31}$$
$$\|x - \eta \nabla c\|^2 \leq (1 - \eta\lambda)^2 \|x\|^2 + \frac{H'\eta^2 + \eta\lambda - 1}{H'} \left(\|g_\|\|^2 + \|g_\perp\|^2\right) \tag{32}$$

Since $\eta < 1$, $H'\eta^2 + \eta\lambda - 1 < H'\eta + \eta\lambda - 1 < 0$. The result follows directly.

■

Lemma 16 *Given a convex function L where ∇L is Lipschitz continuous, define $c(x) = \frac{\lambda}{2}x^2 + L(x)$. If $\eta < \left(\lambda + \|\nabla L\|_{\text{Lip}}\right)^{-1}$, then for all $x \in M$:*

$$d(x - \eta \nabla c(x), x^*) \leq (1 - \eta\lambda) d(x, x^*) \tag{33}$$

Proof Define x^* to be the optimal point, and $f(x) = c(x) - \frac{\lambda}{2}(x - x^*)^2$. Then:

$$f(x) = c(x) - \frac{\lambda}{2}x^2 + \lambda x \cdot x^* - \frac{\lambda}{2}(x^*)^2 \tag{34}$$
$$f(x) = L(x) + \lambda x \cdot x^* - \frac{\lambda}{2}(x^*)^2 \tag{35}$$

For any $x, y \in M$:

$$\nabla f(x) - \nabla f(y) = (\nabla L(x) + \lambda x^*) - (\nabla L(y) + \lambda x^*) \tag{36}$$
$$\nabla f(x) - \nabla f(y) = (\nabla L(x) - \nabla L(y)) \tag{37}$$
$$\|\nabla f(x) - \nabla f(y)\| = \|\nabla L(x) - \nabla L(y)\| \tag{38}$$

Thus, $\|\nabla f\|_{\text{Lip}} = \|\nabla L\|_{\text{Lip}}$. Thus we can apply Lemma 15. ■

Theorem 17 *Given a convex function L where ∇L is Lipschitz continuous, define $c(x) = \frac{\lambda}{2}x^2 + L(x)$. If $\eta < \left(\lambda + \|\nabla L\|_{\text{Lip}}\right)^{-1}$, then for all $x, y \in M$:*

$$d(x - \eta\nabla c(x), y - \eta\nabla c(y)) \leq (1 - \eta\lambda)d(x, y) \tag{39}$$

Proof We prove this by using Lemma 16. In particular, we use a trick insipired by Classical mechanics: instead of studying the dynamics of the update function directly, we change the frame of reference such that one point is constant. This constant point not only does not move, it is also an optimal point in the new frame of reference, so we can use Lemma 16.

Define $g(w) = c(w) - \nabla c(x) \cdot (w - x)$. Note that, for any $y, z \in M$:

$$d(y - \eta\nabla g(y), z - \eta\nabla g(z)) = d(y - \eta\nabla c(y) + \eta\nabla c(x), z - \eta\nabla c(z) + \eta\nabla c(x)) \tag{40}$$

$$d(y - \eta\nabla g(y), z - \eta\nabla g(z)) = \|y - \eta\nabla c(y) + \eta\nabla c(x) - (z - \eta\nabla c(z) + \eta\nabla c(x))\| \tag{41}$$

$$d(y - \eta\nabla g(y), z - \eta\nabla g(z)) = \|y - \eta\nabla c(y) - (z - \eta\nabla c(z))\| \tag{42}$$

$$d(y - \eta\nabla g(y), z - \eta\nabla g(z)) = d(y - \eta\nabla c(y), z - \eta\nabla c(z)) \tag{43}$$

Therefore, g provides a frame of reference where the relative distances between where everything is will be the same as it would be with c. Moreover, note that g is convex, and $\nabla g(x) = 0$. Thus x is the minimizer of g. Moreover, since $g(w) = c(w) - \nabla c(x) \cdot (w - x) = \frac{\lambda}{2}w^2 + L(w) - \nabla c(x) \cdot (w - x)$. If we define $C(w) = L(w) - \nabla c(x) \cdot (w - x)$, then C is convex and $\|\nabla C\|_{\text{Lip}} = \|\nabla L\|_{\text{Lip}}$. Therefore we can apply Lemma 16 with C instead of L, and then we find that $d(y - \eta\nabla g(y), x) \leq (1 - \eta\lambda)d(y, x)$. From Equation (43), $d(y - \eta\nabla c(y), x - \eta\nabla c(x)) \leq (1 - \eta\lambda)d(y, x)$, establishing the theorem. ■

B Proof of Lemma 3

Lemma 3 *If $c^* = \left\|\frac{\partial L(y,\hat{y})}{\partial \hat{y}}\right\|_{\text{Lip}}$ then, for a fixed i, if $\eta \leq (\|x^i\|^2 c^* + \lambda)^{-1}$, the update rule in Equation 271 is a contraction mapping for the Euclidean distance with Lipschitz constant $1 - \eta\lambda$.*

Proof First, let us break down Equation 271. By gathering terms:

$$\phi^i(w) = (1 - \eta\lambda)w - \eta x^i \frac{\partial}{\partial \hat{y}} L(y^i, \hat{y})|_{w \cdot x^i} \tag{44}$$

Define $u : \mathbf{R} \to \mathbf{R}$ to be equal to $u(z) = \frac{\partial}{\partial z} L(y^i, z)$. Because $L(y, \hat{y})$ is convex in $\hat{y}$, $u(z)$ is increasing, and $u(z)$ is Lipschitz continuous with constant c^*.

$$\phi^i(w) = (1 - \eta\lambda)w - \eta u(w \cdot x^i)x^i \tag{45}$$

We break down w into $w_\parallel$ and $w_\perp$, where $w_\perp \cdot x^i = 0$ and $w_\parallel + w_\perp = w$. Thus:

$$\phi^i(w)_\perp = (1 - \eta\lambda)w_\perp \tag{46}$$

$$\phi^i(w)_\parallel = (1 - \eta\lambda)w_\parallel - \eta u(w_\parallel \cdot x^i)x^i \tag{47}$$

Finally, note that $d(w, v) = \sqrt{d^2(w_\parallel, v_\parallel) + d^2(w_\perp, v_\perp)}$.

Note that given any $w_\perp, v_\perp$, $d(\phi^i(w)_\perp, \phi^i(v)_\perp) = (1 - \eta\lambda)d(w_\perp, v_\perp)$. For convergence in the final, "interesting" dimension parallel to x^i, first we observe that if we define $\alpha(w) = x^i \cdot w$, we can represent the update as:

$$\alpha(\phi^i(w)) = (1 - \eta\lambda)\alpha(w) + \eta y^i u(\alpha(w))(x^i \cdot x^i) \tag{48}$$

Define $\beta = \sqrt{x^i \cdot x^i}$. Note that:

$$\alpha(\phi^i(w)) = (1 - \eta\lambda)\alpha(w) + \eta u(\alpha(w))\beta^2 \tag{49}$$

$$d(w_\parallel, v_\parallel) = \frac{1}{\beta}|\alpha(w) - \alpha(v)| \tag{50}$$

$$d(\phi^i(w)_\parallel, \phi^i(v)_\parallel) = \frac{1}{\beta}\left|((1 - \eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1 - \eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2)\right| \tag{51}$$

Without loss of generality, assume that $\alpha(w) \geq \alpha(v)$. Since $\alpha(w) \geq \alpha(v)$, $u(\alpha(w)) \geq u(\alpha(v))$. By Lipschitz continuity:

$$|u(\alpha(w)) - u(\alpha(v))| \leq c^*|\alpha(w) - \alpha(v)| \tag{52}$$
$$u(\alpha(w)) - u(\alpha(v)) \leq c^*(\alpha(w) - \alpha(v)) \tag{53}$$

Rearranging the terms yields:

$$((1-\eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1-\eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2) = ((1-\eta\lambda)(\alpha(w) - \alpha(v)) - \eta\beta^2(u(\alpha(w)) - u(\alpha(v)) \tag{54}$$

Note that $u(\alpha(w)) \geq u(\alpha(v))$, so $\eta\beta^2(u(\alpha(w)) - u(\alpha(v)) \geq 0$, so:

$$((1-\eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1-\eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2) \leq (1-\eta\lambda)(\alpha(w) - \alpha(v)) \tag{55}$$

Finally, since $u(\alpha(w)) - u(\alpha(v)) \leq c^*(\alpha(w) - \alpha(v))$:

$$((1-\eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1-\eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2) \geq ((1-\eta\lambda)(\alpha(w) - \alpha(v)) - \eta\beta^2 c^*(\alpha(w)) - \alpha(v)) = ((1-\eta\lambda - \eta\beta^2 c^*)(\alpha(w) - \alpha(v)) \tag{56}$$

Since we assume in the state of the theorem, $\eta \leq (\beta^2 c^* + \lambda)^{-1}$, it is the case that $(1-\eta\lambda-\eta\beta^2 c^*) \geq 0$, and:

$$((1-\eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1-\eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2) \geq 0 \tag{57}$$

By Equation (55) and Equation (57), it is the case that:

$$|((1-\eta\lambda)\alpha(w) - \eta u(\alpha(w))\beta^2) - ((1-\eta\lambda)\alpha(v) - \eta u(\alpha(v))\beta^2)| \leq (1-\eta\lambda)(\alpha(w) - \alpha(v)) \tag{58}$$

This implies:

$$d(\phi^i(w)_\parallel, \phi^i(v)_\parallel) \leq \frac{1}{\beta}(1-\eta\lambda)(\alpha(w) - \alpha(v)) \tag{59}$$
$$\leq (1-\eta\lambda)\frac{1}{\beta}|\alpha(w) - \alpha(v)| \tag{60}$$
$$\leq (1-\eta\lambda)\frac{1}{\beta}d(w_\parallel, v_\parallel) \tag{61}$$

This establishes that $d(\phi^i(w), \phi^i(v)) \leq (1-\eta\lambda)d(w, v)$.

■

C Wasserstein Metrics and Contraction Mappings

In this section, we prove Lemma 5, Lemma 6, and Corollary 7 from Section 2.2.

Fact 18 *$x^* = \inf_{x \in X} x$ if and only if:*

1. *for all $x \in X$, $x^* \leq x$, and*
2. *for any $\epsilon > 0$, there exits an $x \in X$ such that $x^* + \epsilon > x$.*

Fact 19 *If for all $\epsilon > 0$, $a + \epsilon \geq b$, then $a \geq b$.*

Lemma 5 *For all i, Given a metric space (M, d) and a contraction mapping ϕ on (M, d) with constant c, $\mathbf{p}$ is a contraction mapping on $(P(M, d), W_i)$ with constant c.*

Proof A contraction mapping is continuous and therefore it is a measurable function on the Radon space (which is a Borel space).

Given two distributions X and Y, define $z = W_i(X,Y)$. By Fact 18, for any $\epsilon > 0$, there exists a $\gamma \in \Gamma(X,Y)$ such that $(W_i(X,Y))^i + \epsilon > \int_{x,y} d(x,y)\mathbf{d}^i\gamma(x,y)$. Define γ' such that for all $E, E' \in M$, $\gamma'(E,E') = \gamma(\phi^{-1}(E), \phi^{-1}(E'))$.

Note that $\gamma'(E,M)$=$\gamma(\phi^{-1}(E), M) = X(\phi^{-1}(E)) = \mathbf{p}(X)(E)$, Thus, the marginal distribution of γ is $\mathbf{p}(X)$, and analogously the other marginal distribution of γ is $\mathbf{p}(Y)$. Since ϕ is a contraction with constant c, it is the case that $cd(\phi(x),\phi(y)) \leq d(x,y)$, and

$$(W_i(X,Y))^i + \epsilon > \int_{x,y} \frac{1}{c^i} d^i(\phi(x),\phi(y))\mathbf{d}\gamma(x,y) \tag{62}$$

$$(W_i(X,Y))^i + \epsilon > \frac{1}{c^i}\int_{x,y} d^i(\phi(x),\phi(y))\mathbf{d}\gamma(x,y) \tag{63}$$

By change of variables:

$$(W_i(X,Y))^i + \epsilon > \frac{1}{c^i}\int_{x,y} d^i(x,y)\mathbf{d}\gamma'(x,y) \tag{64}$$

$$(W_i(X,Y))^i + \epsilon > \frac{1}{c^i}(W_i(\mathbf{p}(X),\mathbf{p}(Y)))^i \tag{65}$$

By Fact 19:

$$(W_i(X,Y))^i \geq \frac{1}{c^i}(W_i(\mathbf{p}(X),\mathbf{p}(Y)))^i \tag{66}$$

$$W_i(X,Y) \geq \frac{1}{c}(W_i(\mathbf{p}(X),\mathbf{p}(Y))) \tag{67}$$

Since X and Y are arbitrary, $\mathbf{p}$ is a contraction mapping with metric W_i. ■

Lemma 20 *Given $X^1 \ldots X^m$, $Y^1 \ldots Y^m$ that are probability measures over (M,d), $a_1 \ldots a_m \in \mathbf{R}$, where $\sum_i a_i = 1$ and if for all i, $a_i \geq 0$, and for all i, $W_k(X^i,Y^i)$ is well-defined, then:*

$$W_k\left(\sum_i a_i X^i, \sum_i a_i Y^i\right) \leq \left(\sum_i a_i (W_k(X^i,Y^i))^k\right)^{1/k} \tag{68}$$

Corollary 21 *If for all i, $W_k(X^i,Y^i) \leq d$,then:*

$$W_k\left(\sum_i a_i X^i, \sum_i a_i Y^i\right) \leq d \tag{69}$$

Proof

By Fact 18, for any $\epsilon > 0$, there exists a $\gamma^i \in \Gamma(X^i,Y^i)$ such that:

$$(W_k(X^i,Y^i))^k + \epsilon > \int d^k(x,y)\mathbf{d}\gamma^k(x,y) \tag{70}$$

Note that $\sum_i a_i\gamma^i \in \Gamma(\sum_i a_i X^i, \sum_i a_i Y^i)$, where we consider addition on functions over measureable sets in $(M,d)\times(M,d)$. If we define $\gamma^* = \sum_i a_i\gamma^i$, then:

$$\sum_i a_i \int d^k(x,y)\mathbf{d}\gamma^i(x,y) = \int d^k(x,y)\mathbf{d}\gamma^*(x,y) \tag{71}$$

Therefore:

$$\sum a_i((W_k(X^i,Y^i))^k+\epsilon) > \int d^k(x,y)\mathbf{d}\gamma^*(x,y) \tag{72}$$

$$\epsilon + \sum a_i(W_k(X^i,Y^i))^k > \int d^k(x,y)\mathbf{d}\gamma^*(x,y) \tag{73}$$

$$\tag{74}$$

Because $\gamma^* \in \Gamma(\sum_i a_i X^i, \sum_i a_i Y^i)$:

$$\epsilon + \sum a_i(W_k(X^i,Y^i))^k > \inf_{\gamma\in\Gamma(\sum_i a_i X^i, \sum_i a_i Y^i)} \int d^k(x,y)\mathbf{d}\gamma(x,y) \tag{75}$$

$$\epsilon + \sum a_i(W_k(X^i,Y^i))^k > (W_k(\sum_i a_i X^i, \sum_i a_i Y^i))^k \tag{76}$$

By Fact 19:

$$\sum a_i(W_k(X^i,Y^i))^k \geq (W_k(\sum_i a_i X^i, \sum_i a_i Y^i))^k \tag{77}$$

$$\left(\sum a_i(W_k(X^i,Y^i))^k\right)^{1/k} \geq W_k(\sum_i a_i X^i, \sum_i a_i Y^i) \tag{78}$$

■

Lemma 6 *Given a Radon space (M,d), if $\mathbf{p}_1 \ldots \mathbf{p}_k$ are contraction mappings with constants $c_1 \ldots c_k$ with respect to W_z, and $\sum_i a_i = 1$ where $a_i \geq 0$, then $\mathbf{p} = \sum_{i=1}^k a_i \mathbf{p}_i$ is a contraction mapping with a constant of no more than $(\sum_i a_i (c_i)^z)^{1/z}$.*

Corollary 7 *If for all i, $c_i \leq c$, then $\mathbf{p}$ is a contraction mapping with a constant of no more than c.*

Proof Given an initial measures X, Y, for any i,

$$W_z(\mathbf{p}_i(X), \mathbf{p}_i(Y)) < c_i W_z(X,Y) \tag{79}$$

. Thus, $\mathbf{p}(X) = \sum_{i=1}^k a_i \mathbf{p}_i(X)$ and $\mathbf{p}(Y) = \sum_{i=1}^k a_i \mathbf{p}_i(Y)$, by Lemma 20 it is the case that:

$$W_z(\mathbf{p}(X), \mathbf{p}(Y) \leq \left(\sum_{i=1}^k a_i \left(W_z(\mathbf{p}_i(X), \mathbf{p}_i(Y))\right)^z\right)^{1/z} \tag{80}$$

By Equation 79:

$$W_z(\mathbf{p}(X), \mathbf{p}(Y)) \leq \left(\sum_{i=1}^k a_i \left(c_i W_z(X,Y)\right)^z\right)^{1/z} \tag{81}$$

$$\leq \left(\sum_{i=1}^k a_i \left(c_i W_z(X,Y)\right)^z\right)^{1/z} \tag{82}$$

$$\leq W_z(X,Y)\left(\sum_{i=1}^k a_i \left(c_i\right)^z\right)^{1/z} \tag{83}$$

■

D More Properties of Wasserstein Metrics

D.1 Kantorovich-Rubinstein Theorem

Define $\beta(P, Q)$ to be:

$$\beta(P, Q) = \sup_{f, \|f\|_{\text{Lip}} \leq 1} \left| \int f dP - \int f dQ \right| \tag{84}$$

Where $\|\circ\|_{\text{Lip}}$ is the Lipschitz constant of the function.

Theorem 22 *(Kantorovich-Rubinstein) If (M, d) is a separable metric space then for any two distributions P,Q, we have $W_1(P, Q) = \beta(P, Q)$.*

Corollary 23 *If d is Euclidean distance, $d(\mu_P, \mu_Q) \leq W_1(P, Q)$.*

The following extends one half of Kantorovich-Rubinstein beyond W_1.

Theorem 24 *For any $i \geq 1$, for any f where $\|f\|_{\text{Lip}_i}$ is bounded, for distributions X, Y:*

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] \leq \|f\|_{\text{Lip}_i} (W_i(X, Y))^i . \tag{85}$$

Corollary 25 *Given two distributions X, Y, given any Lipschitz continuous function $c : M \to \mathbf{R}$:*

$$|\mathbf{E}_{x \in X}[c(x)] - \mathbf{E}_{x \in Y}[c(x)]| \leq \|c\|_{\text{Lip}} W_1(X, Y) \tag{86}$$

Proof Choose an arbitrary $i \geq 1$. Choose an f where $\|f\|_{\text{Lip}_i}$ is bounded, and arbitrary distributions X, Y. Choose a joint distribution $\gamma \in (M, d) \times (M, d)$ such that the first marginal of γ is X, and the second marginal of γ is Y. Therefore:

$$\mathbf{E}_{x \in X}[f(x)] = \int f(x) \mathbf{d}\gamma(x, y) \tag{87}$$

$$\mathbf{E}_{y \in Y}[f(y)] = \int f(y) \mathbf{d}\gamma(x, y) \tag{88}$$

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] = \int f(x) \mathbf{d}\gamma(x, y) - \int f(y) \mathbf{d}\gamma(x, y) \tag{89}$$

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] = \int (f(x) - f(y)) \mathbf{d}\gamma(x, y) \tag{90}$$

By the definition of $\|f\|_{\text{Lip}_i}$, $f(x) - f(y) \leq \|f\|_{\text{Lip}_i} d^i(x, y)$:

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] \leq \int \|f\|_{\text{Lip}_i} d^i(x, y) \mathbf{d}\gamma(x, y) \tag{91}$$

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] \leq \|f\|_{\text{Lip}_i} \int d^i(x, y) \mathbf{d}\gamma(x, y) \tag{92}$$

For any $\epsilon > 0$, there exists a γ such that $(W_i(x, y))^i + \epsilon > \int d^i(x, y) \mathbf{d}\gamma(x, y)$. Therefore, for any $\epsilon > 0$:

$$\mathbf{E}_{x \in X}[f(x)] - \mathbf{E}_{y \in Y}[f(y)] \leq \|f\|_{\text{Lip}_i} (W_i(x, y))^i + \epsilon \tag{93}$$

Therefore, if we allow ϵ to approach zero, we prove the theorem. ■

D.2 Wasserstein Distance and Relative Standard Deviation

Before we introduce relative standard deviation, we want to make a few observations about Wasserstein distances and point masses. Given $x \in M$, define $I_x \in P(M,d)$ such that $I_x(E) = 1$ if $x \in E$, and $I_x(E) = 0$ if $x \notin E$. Given $x \in M$ and $Y \in P(M,d)$, define $W_z(x,Y) = W_z(I_x, Y)$. It is the case that:

$$W_z(x,Y) = (\mathbf{E}_{y \in Y}[d^z(x,y)])^{1/i} \tag{94}$$

Lemma 26 *Given $Y \in (M,d)$, $x \in M$, if $\Pr[d(x,y) \leq L] = 1$, then $W_z(x,Y) \leq L$.*

Corollary 27 *For $x, y \in M$, $W_z(x,y) = d(x,y)$.*

Proof Since $\Gamma(I_x, Y)$ is a singleton:

$$W_z(x,Y) = \left(\int d^z(x,y) \mathbf{d}Y(y) \right)^{1/z}. \tag{95}$$

Therefore, we can bound $d^z(x,y)$ by L^z, so:

$$W_z(x,Y) \leq \left(\int L^z \mathbf{d}Y(y) \right)^{1/z} \tag{96}$$

$$W_z(x,Y) \leq (L^z)^{1/z} \tag{97}$$

$$W_z(x,Y) \leq L \tag{98}$$

■

Let us define the **relative standard deviation of X with respect to c** to be:

$$\sigma_X^c = \sqrt{\mathbf{E}[(X-c)^2]}. \tag{99}$$

Define μ_X to be the mean of X. Observe that $\sigma_X = \sigma_X^{\mu_X}$.

Fact 28 *If σ_X^c is finite, then $\sigma_X^c = W_2(I_c, X)$.*

Lemma 29

$$|\sigma_X^c - \sigma_X^{c'}| \leq d(c,c') \tag{100}$$

Proof By the triangle inequality, $W_2(I_c, X) \leq W_2(I_{c'}, X) + W_2(I_c, I_{c'})$. By Fact 28, $\sigma_X^c \leq \sigma_X^{c'} + W_2(I_c, I_{c'})$. By Corollary 27, $\sigma_X^c \leq \sigma_X^{c'} + d(c,c')$. Similarly, one can show $\sigma_X^{c'} \leq \sigma_X^c + d(c,c')$. ■

Lemma 30

$$\sigma_Y^c \leq \sigma_X^c + W_2(X,Y) \tag{101}$$

Proof By the triangle inequality, $W_2(I_c, Y) \leq W_2(I_c, X) + W_2(X,Y)$. The result follows from Fact 28. ■

Theorem 31

$$\sigma_X \leq \sigma_X^c \tag{102}$$

Proof We prove this by considering σ_X^c a function of c, and finding the minimum by checking where the gradient is zero. ■

Theorem 32

$$\sigma_Y \leq \sigma_X + W_2(X, Y) \tag{103}$$

Proof Note that $\sigma_X = \sigma_X^{\mu_X}$. By Lemma 30:

$$\sigma_Y^{\mu_X} \leq \sigma_X^{\mu_X} + W_2(X, Y) \tag{104}$$

By Theorem 31, $\sigma_Y^{\mu_Y} \leq \sigma_Y^{\mu_X}$, proving the result. ■

Theorem 33 *For any d, for any P, Q, if W_i exists, then:*

$$W_i(P, Q) \geq W_1(P, Q) \tag{105}$$

Proof For any $\epsilon > 0$, there exists a $\gamma \in \Gamma(P, Q)$ such that:

$$(W_i(P, Q))^i + \epsilon \geq \int d^i(x, y) \mathbf{d}\gamma(x, y) \tag{106}$$

By Jensen's inequality:

$$\int d^i(x, y) \mathbf{d}\gamma(x, y) \geq \left(\int d(x, y) \mathbf{d}\gamma(x, y) \right)^i \tag{107}$$

Therefore:

$$(W_i(P, Q))^i + \epsilon \geq \left(\int d(x, y) \mathbf{d}\gamma(x, y) \right)^i \tag{108}$$

By definition, $W_1(P, Q) \leq \int d(x, y) \mathbf{d}\gamma(x, y)$, so:

$$(W_i(P, Q))^i + \epsilon \geq (W_1(P, Q))^i \tag{109}$$

Since for any $\epsilon > 0$, this holds, by Fact 19:

$$(W_i(P, Q))^i \geq (W_1(P, Q))^i \tag{110}$$

Since $i \geq 1$, the result follows.

■

Theorem 34 *Suppose that $X^1 \ldots X^k$ are independent and identically distributed random variables over $\mathbf{R}^n$. Then, if $A = \frac{1}{k} \sum_{i=1}^k X^i$, it is the case that:*[1]

$$\mu_A = \mu_{X^1} \tag{111}$$

$$\sigma_A \leq \frac{\sigma_{X^1}}{\sqrt{k}}. \tag{112}$$

Proof

The first is a well known theorem; $\mu_A = \mu_{X^1}$ by linearity of expectation. The second part is one of many direct results of the fact that the variance of two independent variables X and Y is the sum of the variance of the independent variables. ■

[1] Here we mean to indicate the average of the random variables, not the average of their distributions.

D.3 Wasserstein Distance and Cesaro Summability

Theorem 35 *For any Lipschitz continuous function c, for any sequence of distributions $\{D_1, D_2 \ldots\}$ in the Wasserstein metric, if $\lim_{t\to\infty} D_t = D^*$, then:*

$$\lim_{t\to\infty} \mathbf{E}_{x\in D_t}[c(x)] = \mathbf{E}_{x\in D^*}[c(x)] \tag{113}$$

Proof Assume that the Lipschitz constant for c is c^*. By Corollary 25, it is the case that:

$$|\mathbf{E}_{x\in D_t}[c(x)] - \mathbf{E}_{x\in D^*}[c(x)]| \leq c^* W_1(D_t, D^*) \tag{114}$$

We can prove that:

$$\lim_{t\to\infty} |\mathbf{E}_{x\in D_t}[c(x)] - \mathbf{E}_{x\in D^*}[c(x)]| \leq \lim_{t\to\infty} c^* W_1(D_t, D^*) \tag{115}$$

$$\leq c^* \lim_{t\to\infty} W_1(D_t, D^*) \tag{116}$$

$$\leq c^* \times 0 = 0 \tag{117}$$

So, if the distance between the sequence $\{E_{x\in D_t}[c(x)]\}_t$ and the point $\mathbf{E}_{x\in D^*}[c(x)]$ approaches zero, the limit of the sequence is $\mathbf{E}_{x\in D^*}[c(x)]$.

■

Theorem 36 ***(Cesàro Sum)*** *Given a sequence $\{a_1, a_2 \ldots\}$ where $\lim_{t\to\infty} a_t = a^*$, it is the case that:*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} a_t = a^* \tag{118}$$

Proof

For a given $\epsilon > 0$, there exists an t such that for all $t' > t$, $|a_{t'} - a^*| < \frac{\epsilon}{2}$. Define $a_{\mathbf{begin}} = \sum_{t'=1}^{t} a_{t'}$. Then, we know that, for $T > t$:

$$\frac{1}{T} \sum_{t'=1}^{T} a_t = \frac{1}{T} \left(\sum_{t'=1}^{t} a_{t'} + \sum_{t'=t+1}^{T} a_{t'} \right) \tag{119}$$

$$\frac{1}{T} \sum_{t'=1}^{T} a_t = \frac{1}{T} \left(a_{\mathbf{begin}} + \sum_{t'=t+1}^{T} a_{t'} \right) \tag{120}$$

$$\frac{1}{T} \sum_{t'=1}^{T} a_t \leq \frac{1}{T} \left(a_{\mathbf{begin}} + \sum_{t'=t+1}^{T} \left(a^* + \frac{\epsilon}{2} \right) \right) \tag{121}$$

$$\frac{1}{T} \sum_{t'=1}^{T} a_t \leq \frac{1}{T} \left(a_{\mathbf{begin}} + (T-t) \left(a^* + \frac{\epsilon}{2} \right) \right) \tag{122}$$

Note that as $T \to \infty$:

$$\lim_{T\to\infty} \frac{1}{T} \left(a_{\mathbf{begin}} + (T-t) \left(a^* + \frac{\epsilon}{2} \right) \right) = \lim_{T\to\infty} \frac{t}{T} a_{\mathbf{begin}} + \frac{T-t}{T} \left(a^* + \frac{\epsilon}{2} \right) \tag{123}$$

$$= 0 \times a_{\mathbf{begin}} + 1 \times \left(a^* + \frac{\epsilon}{2} \right) \tag{124}$$

$$= a^* + \frac{\epsilon}{2} \tag{125}$$

Therefore, since the upper bound on the limit approaches $a^* + \frac{\epsilon}{2}$, there must exist a T such that for all $T' > T$:

$$\frac{1}{T'+1} \sum_{t=1}^{T'} a_t < a^* + \epsilon \tag{126}$$

Similarly, one can prove that there exists a T'' such that for all $T' > T''$, $\frac{1}{T'+1}\sum_{t=1}^{T'} a_t > a^* - \epsilon$. Therefore, the series converges.

■

Theorem 37 *For any Lipschitz continuous function c, for any sequence of distributions $\{D_1, D_2 \ldots\}$ in the Wasserstein metric, if $\lim_{t\to\infty} D_t = D^*$, then:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{x\in D_t}[c(x)] = \mathbf{E}_{x\in D^*}[c(x)] \tag{127}$$

Proof This is a direct result of Theorem 35 and Theorem 36. ■

E Basic Properties of Stochastic Gradient Descent on SVMs

$$\nabla c^i(w) = \lambda w + \frac{\partial}{\partial \hat{y}} L(y^i, \hat{y})|_{w^i \cdot x^i} x^i \tag{128}$$

Define f such that:

$$f^i(w) = L(y^i, w^i \cdot x^i) \tag{129}$$

We assume that for all i, for all w, $\left\|\nabla f^i(w)\right\| \leq G$. Also, define:

$$f(w) = \frac{1}{m}\sum_{i=1}^{m} f^i(w) \tag{130}$$

In order to understand the stochastic process, we need to understand the batch update. The expected stochastic update is the batch update. Define g_w to be the expected gradient at w, and $c(w)$ to be the expected cost at w.

$$c(w) = \frac{\lambda}{2}w^2 + f(w) \tag{131}$$

Theorem 38 *The expected gradient is the gradient of the expected cost.*

Proof This follows directly from the linearity of the gradient operator and the linearity of expectation. ■

The following well-known theorem establishes that c is a strongly convex function.

Theorem 39 *For any w, w':*

$$c(w') \geq \frac{\lambda}{2}(w' - w)^2 + g_w \cdot (w' - w) + c(w) \tag{132}$$

Proof

$\frac{\lambda}{2}w^2$ is a λ- strongly convex function, and $f^i(w)$ is a convex function, so therefore $c(w)$ is a λ-strongly convex function. Or, to be more thorough, because f is convex:

$$f(w') - f(w) \geq \nabla f(w) \cdot (w' - w). \tag{133}$$

Define $h(w) = \frac{\lambda}{2}w^2$. Observe that:

$$h(w') - h(w) = \frac{\lambda}{2}(w')^2 - \frac{\lambda}{2}w^2 \tag{134}$$

$$h(w') - h(w) = \frac{\lambda}{2}(w')^2 - \frac{\lambda}{2}w^2 - \lambda w \cdot (w' - w) + \lambda w \cdot (w' - w) \tag{135}$$

$$h(w') - h(w) = \frac{\lambda}{2}(w')^2 - \frac{\lambda}{2}w^2 - \lambda w \cdot w' + \lambda w^2 + \lambda w \cdot (w' - w) \tag{136}$$

$$h(w') - h(w) = \frac{\lambda}{2}(w')^2 + \frac{\lambda}{2}w^2 - \lambda w \cdot w' + \lambda w \cdot (w' - w) \tag{137}$$

$$h(w') - h(w) = \frac{\lambda}{2}(w')^2 + \frac{\lambda}{2}w^2 - \lambda w \cdot w' + \nabla h(w) \cdot (w' - w) \tag{138}$$

$$h(w') - h(w) = \frac{\lambda}{2}(w' - w)^2 + \nabla h(w) \cdot (w' - w) \tag{139}$$

Since $c(w) = h(w) + f(w)$:

$$c(w') - c(w) \geq \frac{\lambda}{2}(w' - w)^2 + \nabla h(w) \cdot (w' - w) + \nabla f(w) \cdot (w' - w) \tag{140}$$

$$c(w') - c(w) \geq \frac{\lambda}{2}(w' - w)^2 + \nabla c(w) \cdot (w' - w) \tag{141}$$

■

Theorem 40

$$\|w^*\| \leq \frac{G}{\lambda}.$$

Proof Note that $\nabla c(w^*) = 0$. So:

$$0 = \nabla c(w^*) \tag{142}$$

$$0 = \nabla \left(\frac{\lambda}{2}(w^*)^2 + f(w^*) \right) \tag{143}$$

$$0 = \lambda w^* + \nabla f(w^*) - \lambda w^* \tag{144}$$

$$= \nabla f(w^*) \tag{145}$$

Since $\|\nabla f(w^*)\| \leq G$, it is the case that:

$$\|-\lambda w^*\| \leq G \tag{146}$$

$$\lambda \|w^*\| \leq G \tag{147}$$

$$\|w^*\| \leq \frac{G}{\lambda} \tag{148}$$

■

Theorem 41 *For any w, if w^* is the optimal point:*

$$\lambda(w^* - w)^2 \leq g_w \cdot (w - w^*) \tag{149}$$

Proof By Theorem 39:

$$c(w^*) \geq \frac{\lambda}{2}(w^* - w)^2 + g_w \cdot (w^* - w) + c(w) \tag{150}$$

$$c(w^*) - c(w) \geq \frac{\lambda}{2}(w^* - w)^2 + g_w \cdot (w^* - w) \tag{151}$$

$$c(w) - c(w^*) \leq -\frac{\lambda}{2}(w^* - w)^2 + g_w \cdot (w - w^*) \tag{152}$$

Since w^* is optimal, $\nabla c(w^*) = 0$, implying:

$$c(w) \geq \frac{\lambda}{2}(w^* - w)^2 + 0 \cdot (w - w^*) + c(w^*) \tag{153}$$

$$c(w) - c(w^*) \geq \frac{\lambda}{2}(w^* - w)^2 \tag{154}$$

Combining Equation 152 and Equation 154:

$$\frac{\lambda}{2}(w^* - w)^2 \leq -\frac{\lambda}{2}(w^* - w)^2 + g_w \cdot (w - w^*) \tag{155}$$

$$\lambda(w^* - w)^2 \leq g_w \cdot (w - w^*) \tag{156}$$

■

Theorem 42 *For any w:*

$$\left\|\nabla c^i - \lambda(w - w^*)\right\| \leq 2G \tag{157}$$

Proof First, observe that:

$$\nabla c^i(w) = \lambda w + \nabla f^i(w) \tag{158}$$

$$\nabla c^i(w) - \lambda w \leq \nabla f^i(w) \tag{159}$$

$$\left\|\nabla c^i(w) - \lambda w\right\| \leq G \tag{160}$$

Also, $\|w^*\| \leq \frac{G}{\lambda}$, implying $\|\lambda w^*\| \leq G$. Thus, the triangle inequality yields:

$$\left\|(\nabla c^i(w) - \lambda w) + (\lambda w^*)\right\| \leq 2G \tag{161}$$

$$\left\|\nabla c^i(w) - \lambda(w - w^*)\right\| \leq 2G \tag{162}$$

■

Thus, minus a contraction ratio, the magnitude of the gradient is bounded. Moreover, in expectation it is not moving away from the optimal point. These two facts will help us to bound the expected mean and expected squared distance from optimal.

Theorem 43 *For any w, if w^* is the optimal point, and $\eta \in (0, 1)$:*

$$((w - \eta g_w) - w^*) \cdot (w - w^*) \leq (1 - \eta\lambda)(w - w^*)^2 \tag{163}$$

Proof

From Theorem 41,

$$\lambda(w^* - w)^2 \leq g_w \cdot (w - w^*). \tag{164}$$

Multiplying both sides by η:

$$\eta\lambda(w^* - w)^2 \leq \eta g_w \cdot (w - w^*) \tag{165}$$

$$-\eta g_w \cdot (w - w^*) \leq -\eta\lambda(w^* - w)^2 \tag{166}$$

Adding $(w - w^*) \cdot (w - w^*)$ to both sides yields the result. ■

Theorem 44 *If w_t is a state of the stochastic gradient descent algorithm, $w_0 = 0$, $\lambda \leq 1$, and $0 \leq \eta \leq \frac{1}{\lambda}$, then:*

$$\|w_t\| \leq \frac{G}{\lambda} \tag{167}$$

Corollary 45

$$\left\|\nabla c^i(w_t)\right\| \leq 2G \tag{168}$$

Proof First, observe that $\|w_0\| \leq \frac{G}{\lambda}$. We prove the theorem via induction on t. Assume that the condition holds for $t-1$, i.e.that $\|w_{t-1}\| \leq \frac{G}{\lambda}$. Then, w_t is, for some i:

$$w_t \leq w_{t-1}(1-\eta\lambda) - \eta\nabla f^i(w_t) \tag{169}$$

$$\|w_t\| \leq |1-\eta\lambda| \|w_{t-1}\| + |\eta| \left\|\nabla f^i(w_t)\right\| \tag{170}$$

Since $\|w_{t-1}\| \leq \frac{G}{\lambda}$ and $\left\|\nabla f^i(w_t)\right\| \leq G$, then:

$$\|w_t\| \leq |1-\eta\lambda|\frac{G}{\lambda} + |\eta|G \tag{171}$$

Since $\eta \geq 0$ and $1-\eta\lambda \geq 0$:

$$\|w_t\| \leq (1-\eta\lambda)\frac{G}{\lambda} + \eta G \tag{172}$$

$$\|w_t\| \leq \frac{G}{\lambda} \tag{173}$$

■

F Proof of Theorem 8: SGD is a Contraction Mapping

Theorem 8 *For any positive integer z, if $\eta \leq \eta^*$, then $\mathbf{p}^*$ is a contraction mapping on (M, W_z) with contraction rate $(1-\eta\lambda)$. Therefore, there exists a unique D_η^* such that $\mathbf{p}^*(D_\eta^*) = D_\eta^*$. Moreover, if $w_0 = 0$ with probability 1, then $W_z(D_\eta^0, D_\eta^*) = \frac{G}{\lambda}$, and $W_z(D_\eta^T, D_\eta^*) \leq \frac{G}{\lambda}(1-\eta\lambda)^T$.*

Proof The contraction rate $(1-\eta\lambda)$ can be proven by applying Lemma 3, Lemma 5, and Corollary 6. By Theorem 44, $\|w_t\| \leq \frac{G}{\lambda}$. Therefore, for any $w \in D_\eta^*$, $\|w\| \leq \frac{G}{\lambda}$. Since $D_\eta^0 = I_{w_0}$, it is the case that $W_z(D_\eta^0, D_\eta^*) = W_z(0, D_\eta^*)$. By Lemma 26, $W_z(D_\eta^0, D_\eta^*) \leq \frac{G}{\lambda}$. By applying the first half of the theorem and Corollary 2, $W_z(D_\eta^T, D_\eta^*) \leq \frac{G}{\lambda}(1-\eta\lambda)^T$. ■

G Proof of Theorem 9: Bounding the Error of the Mean

Define $\frac{D}{2}$ to be a bound on the distance the gradient descent algorithm can be from the origin. Therefore, we can use the algorithm and analysis from [11], where we say D is the diameter of the space, and M is the maximum gradient in that space. However, we will use a constant learning rate.

Theorem 46 *Given a sequence $\{c_t\}$ of convex cost functions, a domain F that contains all vectors of the stochastic gradient descent algorithm, a bound M on the norm of the gradients of c_t in F. The regret of stochastic gradient descent algorithm after T time steps is:*

$$R_T = \underset{w^* \in F}{\operatorname{argmax}} \sum_{t=1}^{T} (c_t(w_t) - c_t(w^*)) \leq \frac{T\eta M^2}{2} + \frac{D^2}{2\eta} \tag{174}$$

Proof

We prove this via a potential $\Phi_t = \frac{1}{2\eta}(w_{t+1} - w^*)^2$. First observe that, because c_t is convex:

$$c_t(w^*) \geq (w^* - w_t)\nabla c_t(w_t) + c_t(w_t) \tag{175}$$

$$c_t(w_t) - c_t(w^*) \leq (w_t - w^*)\nabla c_t(w_t) \tag{176}$$

$$R_t - R_{t-1} \leq (w_t - w^*)\nabla c_t(w_t) \tag{177}$$

Also, note that:

$$\Phi_t - \Phi_{t-1} = \frac{1}{2\eta}(w_t - \eta\nabla c_t(w_t) - w^*)^2 - \frac{1}{2\eta}(w_t - w^*)^2 \tag{178}$$

$$\Phi_t - \Phi_{t-1} = -(w_t - w^*)\nabla c_t(w_t) + \frac{\eta}{2}(\nabla c_t(w_t))^2 \tag{179}$$

Adding Equation (177) and Equation (179) then cancelling the $(w_t - w^*)\nabla c_t(w_t)$ terms yields:

$$(R_t - R_{t-1}) + (\Phi_t - \Phi_{t-1}) \leq \frac{\eta}{2}(\nabla c_t(w_t))^2 \tag{180}$$

Summing over all t:

$$\sum_{t=1}^{T}((R_t - R_{t-1}) + (\Phi_t - \Phi_{t-1})) \leq \sum_{t=1}^{T}\frac{\eta}{2}(\nabla c_t(w_t))^2 \tag{181}$$

$$R_T - R_0 \leq \sum_{t=1}^{T}\frac{\eta}{2}(\nabla c_t(w_t))^2 + \Phi_0 - \Phi_T \tag{182}$$

By definition, $R_0 = 0$, and $\Phi_T > 0$, so:

$$R_T \leq \sum_{t=1}^{T}\frac{\eta}{2}(\nabla c_t(w_t))^2 + \Phi_0 \tag{183}$$

$$R_T \leq \sum_{t=1}^{T}\frac{\eta}{2}(\nabla c_t(w_t))^2 + \frac{1}{2\eta}(w_1 - w^*)^2 \tag{184}$$

The distance is bounded by D, and the gradient is bounded by M, so:

$$R_T \leq \frac{T\eta M^2}{2} + \frac{D^2}{2\eta} \tag{185}$$

■

Theorem 47 *Given $c_1 \dots c_m$, if for every $t \in \{1 \dots T\}$, i_t is chosen uniformly at random from 1 to m, then:*

$$\min_{w\in F}\mathbf{E}\left[\sum_{t=1}^{T}c_{i_t}(w)\right] \geq \mathbf{E}\left[\min_{w\in F}\sum_{t=1}^{T}c_{i_t}(w)\right] \tag{186}$$

Proof Observe that, by definition:

$$\mathbf{E}\left[\min_{w\in F}\sum_{t=1}^{T}c_{i_t}(w)\right] = \frac{1}{m^T}\sum_{i_1\dots i_T\in\{1\dots m\}}\min_{w\in F}\sum_{t=1}^{T}c_{i_t}(w) \tag{187}$$

$$\leq \min_{w\in F}\frac{1}{m^T}\sum_{i_1\dots i_T\in\{1\dots m\}}\sum_{t=1}^{T}c_{i_t}(w) \tag{188}$$

$$\leq \min_{w\in F}\mathbf{E}\left[\sum_{t=1}^{T}c_{i_t}(w)\right] \tag{189}$$

■

Theorem 48

$$\lim_{T\to\infty}\frac{1}{T}\mathbf{E}[R_T] \geq \mathbf{E}_{w\in D_\eta^*}[c(w)] - \min_{w\in F}c(w). \tag{190}$$

Proof

This proof follows the technique of many reductions establishing that batch learning can be reduced to online learning [5, 4], but taken to the asymptotic limit. First, observe that

$$\min_{w \in F} \mathbf{E}\left[\sum_{t=1}^{T} c_{i_t}(w)\right] \geq \mathbf{E}\left[\min_{w \in F} \sum_{t=1}^{T} c_{i_t}(w)\right], \tag{191}$$

because it is easier to minimize the utility after the costs are selected. Applying this, the linearity of expectation, and the definitions of c and D_η^t one obtains:

$$\mathbf{E}[R_T] \geq \sum_{t=1}^{T} \mathbf{E}_{w \in D_\eta^t}\left[c(w)\right] - T \min_{w \in F} c(w). \tag{192}$$

Taking the Cesàro limit of both sides yields:

$$\lim_{T \to \infty} \frac{1}{T} \mathbf{E}[R_T] \geq \lim_{T \to \infty} \frac{1}{T} \left(\sum_{t=1}^{T} \mathbf{E}_{w \in D_\eta^t}\left[c(w)\right] - T \min_{w \in F} c(w) \right). \tag{193}$$

The result follows from Theorem 8 and Theorem 37:

■

Theorem 49 *If D_η^* is the stationary distribution of the stochastic update with learning rate η, then:*

$$\frac{\eta M^2}{2} \geq \mathbf{E}_{w \in D_\eta^*}[c(w)] - \min_{w \in F} c(w) \tag{194}$$

Proof From Theorem 48, we know:

$$\lim_{T \to \infty} \frac{1}{T} \mathbf{E}[R_T] \geq \mathbf{E}_{w \in D_\eta^*}\left[c(w)\right] - \min_{w \in F} c(w). \tag{195}$$

Applying Theorem 46:

$$\lim_{T \to \infty} \frac{1}{T} \left(\frac{T \eta M^2}{2} + \frac{D^2}{2\eta} \right) \geq \mathbf{E}_{w \in D_\eta^*}\left[c(w)\right] - \min_{w \in F} c(w). \tag{196}$$

Taking the limit on the left-hand side yields the result. ■

Theorem 50 $c(\mathbf{E}_{w \in D_\eta^*}[w]) - \min_{w \in F} c(w) \leq \frac{\eta M^2}{2}$.

Proof By Theorem 49, $\frac{\eta M^2}{2} \geq \mathbf{E}_{w \in D_\eta^*}[c(w)] - \min_{w \in F} c(w)$. Since c is convex, by Jensen's inequality, the cost of the mean is less than or equal to the mean of the cost, formally $\mathbf{E}_{w \in D_\eta^*}[c(w)] \geq c(\mathbf{E}_{w \in D_\eta^*}[w])$, and the result follows by substitution. ■

Theorem 9 $c(\mathbf{E}_{w \in D_\eta^*}[w]) - \min_{w \in \mathbf{R}^n} c(w) \leq 2\eta G^2$.

This is obtained by applying Theorem 45, and substituting $2G$ for M.

H Generalizing Reinforcement Learning

In order to make this theorem work, we have to push the limits of reinforcement learning. In particular, we have to show that some (but not all) of reinforcement learning works if actions can be any subset of the discrete distributions over the next state. In general, the distribution over the next action is rarely restricted in reinforcement learning. In particular, the theory of discounted reinforcement learning works well on almost any space of policies, but we only show infinite horizon average reward reinforcement learning works when the function is a contraction.

If (M, d) is a Radon space, a probability measure $\rho \in P(M, d)$ is **discrete** if there exists a countable set $C \subseteq S$ such that $\rho(C) = 1$. Importantly, if a function $R : M \to \mathbf{R}$ is a bounded (not necessarily continuous) function, then $\mathbf{E}_{x \in \rho}[R(x)]$ is well-defined. We will denote the set of discrete distributions as $D(M, d) \subseteq P(M, d)$.

Given a Radon space (S, d), define S to be the set of states. Define the actions $A = D(S, d)$ to be the set of discrete distributions over S. For every $w \in S$, define $A(w) \subseteq A$ to be the actions available in state w.

We define a policy as a function $\sigma : S \to A$ where $\sigma(w) \in A(w)$. Then, we can write a transformation $T_\sigma : D(S, d) \to D(S, d)$ such that for any measureable set E, $T_\sigma(\rho)(E)$ is the probability that $w' \in E$, given w' is drawn from $\sigma(w)$ where w is drawn from ρ. Therefore:

$$T_\sigma(\rho)(E) = \mathbf{E}_{w \in \rho}[\sigma(w)(E)] \tag{197}$$

Define $r_0(w, \sigma) = R(w)$, and for $t \geq 1$:

$$r_t(w, \sigma) = \mathbf{E}_{w' \in T_\sigma^t(w)}[R(w')] \tag{198}$$

Importantly, $r_t(w, \sigma) \in [a, b]$. Now, we can define the discounted utility:

$$V_{\sigma,\gamma}^T(w) = \sum_{t=0}^{T} \gamma^t r_t(w, \sigma) \tag{199}$$

Theorem 51 *The sequence $V_{\sigma,\gamma}^1(w), V_{\sigma,\gamma}^2(w), V_{\sigma,\gamma}^3(w)$ converges.*

Proof Since $r_t \in [a, b]$, then for any t, $\gamma^t r_t(w, \sigma) \leq \gamma^t b$. For any T, T' where $T' > T$:

$$V_{\sigma,\gamma}^{T'}(w) - V_{\sigma,\gamma}^T(w) = \sum_{t=T+1}^{T'} \gamma^t r_t(w, \sigma) \tag{200}$$

$$\leq b \frac{\gamma^{T+1} - \gamma^{T'+1}}{1 - \gamma} \tag{201}$$

$$\leq b \frac{\gamma^{T+1}}{1 - \gamma} \tag{202}$$

Similarly, $V_{\sigma,\gamma}^T(w) - V_{\sigma,\gamma}^{T'}(w) \leq -a \frac{\gamma^{T+1}}{1-\gamma}$

Thus, for a given T, for all $T', T'' > T$, $|V_{\sigma,\gamma}^{T''}(w) - V_{\sigma,\gamma}^{T'}(w)| < \max(-a, b) \frac{\gamma^{T+1}}{1-\gamma}$.

Therefore, for any $\epsilon > 0$, there exists a T such that for all $T', T'' > T$ where $|V_{\sigma,\gamma}^{T''}(w) - V_{\sigma,\gamma}^{T'}(w)| < \epsilon$. Therefore, the sequence is a Cauchy sequence, and has a limit since the real numbers are complete. ■

Therefore, we can define:

$$V_{\sigma,\gamma}(w) = \sum_{t=0}^{\infty} \gamma^t r_t(w, \sigma) \tag{203}$$

Note that the limit is well-defined, because R is bounded over S. Also, we can define:

$$\bar{V}_{\sigma,T}(w) = \frac{1}{T+1}\sum_{t=0}^{T} r_t(\sigma, w) \tag{204}$$

Consider W_1 to be the Wasserstein metric on $P(S, d)$.

Theorem 52 *If T_σ is a contraction operator on $(P(S,d), W_1)$, and R is Lipshitz continuous on S, then $r_0(\sigma, w), r_1(\sigma, w), r_2(\sigma, w) \ldots$ converges.*

Proof By Theorem 1, there exists a D^* such that for all w, $\lim_{t\to\infty} T_\sigma^t(w) = D^*$. Since $r_t(\sigma, w) = \mathbf{E}_{w' \in T_\sigma^t(w)}[R(w)]$, by Theorem 35, this sequence must have a limit. ■

Theorem 53 *If T_σ is a contraction operator, and R is Lipschitz continuous, then $\bar{V}_{\sigma,1}(w), \bar{V}_{\sigma,2}(w), \ldots$ converges to $\lim_{t\to\infty} r_t(\sigma, w)$.*

Proof From Theorem 52, we know there exists an r^* such that $\lim_{t\to\infty} r_t(\sigma, w) = r^*$. The result follows from Theorem 36. ■

If T_σ is a contraction mapping, and R is Lipschitz continuous, we can define:

$$\bar{V}_\sigma(w) = \lim_{T\to\infty} \bar{V}_{\sigma,T}(w) \tag{205}$$

Theorem 54 *If T_σ is a contraction mapping, and R is Lipschitz continuous, then:*

$$\bar{V}_\sigma(w) = \lim_{\gamma\to 1^-} (1-\gamma) V_{\sigma,\gamma}(w) \tag{206}$$

Proof From Theorem 52, we know there exists an r^* such that $\bar{V}_\sigma(w) = \lim_{t\to\infty} r_t(\sigma, w) = r^*$. We can also show that $\lim_{\gamma\to 1^-}(1-\gamma)V_{\sigma,\gamma}(w) = r^*$.

We will prove that for a given $\epsilon > 0$, there exists a γ such that $|(1-\gamma)V_{\sigma,\gamma}(w) - r^*| < \epsilon$. For $\frac{\epsilon}{2}$, there exists a t such that for all $t' > t$, $|r_{t'}(\sigma, w) - r^*| < \frac{\epsilon}{2}$. Thus,

$$(1-\gamma)V_{\sigma,\gamma}(w) = (1-\gamma)\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'}(\sigma, w) \tag{207}$$

$$(1-\gamma)V_{\sigma,\gamma}(w) = (1-\gamma)\sum_{t'=0}^{t} \gamma^{t'} r_{t'}(\sigma, w) + (1-\gamma)\sum_{t'=t+1}^{\infty} \gamma^{t'} r_{t'}(\sigma, w) \tag{208}$$

$$(1-\gamma)V_{\sigma,\gamma}(w) \geq (1-\gamma)\sum_{t'=0}^{t} \gamma^{t'} a + (1-\gamma)\sum_{t'=t+1}^{\infty} (r^* - \frac{\epsilon}{2}) \tag{209}$$

$$\tag{210}$$

Since $r^* = (1-\gamma)\sum_{t'=0}^{\infty} \gamma^{t'} r^*$:

$$r^* - (1-\gamma)V_{\sigma,\gamma}(w) \leq (1-\gamma)\sum_{t'=0}^{t} \gamma^{t'}(r^* - a) + (1-\gamma)\sum_{t'=t+1}^{\infty} \frac{\epsilon}{2} \tag{211}$$

$$r^* - (1-\gamma)V_{\sigma,\gamma}(w) \leq (1-\gamma)\frac{1-\gamma^{t+1}}{1-\gamma}(r^* - a) + (1-\gamma)\frac{\gamma^{t+1}}{1-\gamma}\frac{\epsilon}{2} \tag{212}$$

$$r^* - (1-\gamma)V_{\sigma,\gamma}(w)(1-\gamma^{t+1})(r^* - a) + \gamma^{t+1}\frac{\epsilon}{2} \tag{213}$$

$$\tag{214}$$

Note that $\lim_{\gamma \to 1^-}(1-\gamma^{t+1}) = 0$, and $\lim_{\gamma \to 1^-}\gamma^{t+1} = 1$, so:

$$\lim_{\gamma \to 1^-}(1-\gamma^{t+1})(r^* - a) + \gamma^{t+1}\frac{\epsilon}{2} = \frac{\epsilon}{2} \tag{215}$$

Therefore, there exists a $\gamma < 1$ such that for all $\gamma' \in (\gamma, 1)$, $r^* - (1-\gamma')V_{\sigma,\gamma'}(w) < \epsilon$. Similarly, one can prove there exists a $\gamma'' < 1$ such that for all $\gamma' \in (\gamma'', 1)$, $(1-\gamma')V_{\sigma,\gamma'}(w) - r^* < \epsilon$. Thus,$\lim_{\gamma \to 1^-}(1-\gamma)V_{\sigma,\gamma}(w) = r^*$.

■

So, the general view is that for σ which result in T being a contraction mapping and R being a reward function, all the natural aspects of value functions hold. However, for *any* σ and for any bounded reward R, the discounted reward is well-defined. What we will do is now bound the discounted reward using an equation very similar to the Bellman equation.

Theorem 55 *For all $w \in S$:*

$$V_{\sigma,\gamma}(w) = R(w) + \gamma \mathbf{E}_{w' \in T_\sigma(w)}\left[V_{\sigma,\gamma}(w')\right] \tag{216}$$

Proof By definition,

$$V_{\sigma,\gamma}(w) = \sum_{t=0}^{\infty}\gamma^t \mathbf{E}_{w' \in T_\sigma^t(w)}[R(w')] \tag{217}$$

$$V_{\sigma,\gamma}(w) = R(w) + \sum_{t=1}^{\infty}\gamma^t \mathbf{E}_{w' \in T_\sigma^t(w)}[R(w')] \tag{218}$$

Note that for any $t \geq 1$, $T_\sigma^t(w) = T_\sigma^{t-1}(T_\sigma(w))$, so:

$$\mathbf{E}_{w' \in T_\sigma^t(w)}[R(w')] = \mathbf{E}_{w' \in T_\sigma(w)}[\mathbf{E}_{w'' \in T^{t-1}(w')}[R(w'')]] \tag{219}$$

$$\mathbf{E}_{w' \in T_\sigma^t(w)}[R(w')] = \mathbf{E}_{w' \in T_\sigma(w)}[r_{t-1}(\sigma, w')] \tag{220}$$

Applying this to the equation above:

$$V_{\sigma,\gamma}(w) = R(w) + \sum_{t=1}^{\infty}\gamma^t \mathbf{E}_{w' \in T_\sigma(w)}[r_{t-1}(\sigma, w')] \tag{221}$$

$$V_{\sigma,\gamma}(w) = R(w) + \gamma\sum_{t=1}^{\infty}\gamma^{t-1} \mathbf{E}_{w' \in T_\sigma(w)}[r_{t-1}(\sigma, w')] \tag{222}$$

$$V_{\sigma,\gamma}(w) = R(w) + \gamma\sum_{t=0}^{\infty}\gamma^{t} \mathbf{E}_{w' \in T_\sigma(w)}[r_{t}(\sigma, w')] \tag{223}$$

By linearity of expectation:

$$V_{\sigma,\gamma}(w) = R(w) + \gamma \mathbf{E}_{w' \in T_\sigma(w)}[\sum_{t=0}^{\infty}\gamma^t r_t(\sigma, w')] \tag{224}$$

$$V_{\sigma,\gamma}(w) = R(w) + \gamma \mathbf{E}_{w' \in T_\sigma(w)}[V_{\sigma,\gamma}(w)] \tag{225}$$

■

The space of value functions for the discount factor γ is $\mathcal{V} = [\frac{a}{1-\gamma}, \frac{b}{1-\gamma}]^S$. For $V \in \mathcal{V}$, for $a \in A$, we define $V(a) = \mathbf{E}_{x \in a}[V(a)]$. We define the **supremum Bellman operator** $\mathbf{V}_{\sup} : \mathcal{V} \to \mathcal{V}$ such that for all $V \in \mathcal{V}$, for all $w \in S$:

$$\mathbf{V}_{\sup}(V)(w) = R(w) + \gamma \sup_{a \in A(w)} V(a) \tag{226}$$

Define $\mathbf{V}_{\sup}^t$ to be t operations of $\mathbf{V}_{\sup}$.

Define the metric $d_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbf{R}$ such that $d_{\mathcal{V}}(V, V') = \sup_{w \in S} |V(w) - V'(w)|$.

Fact 56 *For any discrete distribution $X \in D(S, d)$, for any $V, V' \in \mathcal{V}$, $\mathbf{E}_{x \in X}[V'(x)] \geq \mathbf{E}_{x \in X}[V(x)] - d_{\mathcal{V}}(V, V')$.*

Theorem 57 *$\mathbf{V}_{\sup}$ is a contraction mapping under the metric $d_{\mathcal{V}}$.*

Proof Given any $V, V' \in \mathcal{V}$, for a particular $w \in S$, since $\mathbf{V}_{\sup}(V)(w) = R(w) + \sup_{a \in A(w)} V(a)$:

$$|\mathbf{V}_{\sup}(V)(w) - \mathbf{V}_{\sup}(V')(w)| = \left| \sup_{a \in A(w)} V(a) - \sup_{a' \in A(w)} V'(a') \right| \tag{227}$$

Without loss of generality, $\sup_{a \in A(w)} V(a) \geq \sup_{a \in A(w)} V'(a)$. Therefore, for any $\epsilon > 0$, there exists a $a' \in A(w)$ such that $V(a') > \sup_{a \in A(w)} V(a) - \epsilon$. By Fact 56, $V'(a') \geq V(a') - d_{\mathcal{V}}(V, V')$, and $V(a') - d_{\mathcal{V}}(V, V') > \sup_{a \in A(w)} V(a) - \epsilon - d_{\mathcal{V}}(V, V')$. This implies $\sup_{a \in A(w)} V'(a) \geq V(a) - d_{\mathcal{V}}(V, V')$. Therefore, $\mathbf{V}_{\sup}(V)(w) - \mathbf{V'}_{\sup}(V)(w) \leq \gamma d_{\mathcal{V}}(V, V')$, and $\mathbf{V}_{\sup}(V)(w) - \mathbf{V}_{\sup}(V')(w) \geq 0$. Therefore, for all w:

$$|\mathbf{V}_{\sup}(V)(w) - \mathbf{V}_{\sup}(V')(w)| \leq \gamma d_{\mathcal{V}}(V, V'), \tag{228}$$

which establishes that $\mathbf{V}_{\sup}$ is a contraction mapping. ■

Under the supremum norm, $\mathcal{V}$ is a complete space, implying that $\mathbf{V}_{\sup}$ as a contraction mapping has a unique fixed point by Banach's fixed point theorem. We call the fixed point V^*.

For $V, V' \in \mathcal{V}$, we say $V \succeq V'$ if for all $w \in S$, $V(w) \geq V'(w)$.

Theorem 58 *If $V \succeq V'$, then $\mathbf{V}_{\sup}(V) \succeq \mathbf{V}_{\sup}(V')$.*

Proof We prove this by contradiction. In particular we assume that there exists a $w \in S$ where $\mathbf{V}_{\sup}(V)(w) < \mathbf{V}_{\sup}(V')(w)$. This would imply:

$$\sup_{a \in A(w)} \mathbf{E}_{x \in a}[V(x)] < \sup_{a \in A(w)} \mathbf{E}_{x \in a}[V'(x)] \tag{229}$$

This would imply that there exists an a such that $\mathbf{E}_{x \in a}[V'(x)] > \sup_{a' \in A(w)} \mathbf{E}_{x \in a'}[V(x)] \geq \mathbf{E}_{x \in a}[V(x)]$. However, since $a \in A(w)$ is a discrete distribution, if $V(a) < V'(a)$, there must be a point where $V(w') < V'(w')$, a contradiction. ■

Lemma 59 *If $\mathbf{V}_{\sup}(V) \succeq V$, then for all t, $\mathbf{V}_{\sup}^t(V) \succeq \mathbf{V}_{\sup}^{t-1}(V)$.*

Proof We prove this by induction on t. It holds for $t = 1$, based on the assumptions in the lemma. If we assume it holds for t, then we need to prove it holds for $t + 1$. By Theorem 58, since $\mathbf{V}_{\sup}^{t-1}(V) \succeq \mathbf{V}_{\sup}^{t-2}(V)$, then $\mathbf{V}_{\sup}(\mathbf{V}_{\sup}^{t-1}(V)) \succeq \mathbf{V}_{\sup}(\mathbf{V}_{\sup}^{t-2}(V))$. Of course, this proves our inductive hypothesis. ■

Lemma 60 *If $\mathbf{V}_{\sup}(V) \succeq V$, then for all t, $\mathbf{V}_{\sup}^t(V) \succeq V$, and therefore $V^* \succeq V$.*

Proof Again we prove this by induction on t, and the base case where $t = 1$ is given in the lemma. Assume that this holds for $t - 1$, in other words, $\mathbf{V}_{\sup}^{t-1}(V) \succeq V$. By Lemma 59, $\mathbf{V}_{\sup}^t(V) \succeq \mathbf{V}_{\sup}^{t-1}(V)$, so by transitivity, $\mathbf{V}_{\sup}^t(V) \succeq V$. ■

Theorem 61 *For any σ: For any V such that, for all $w \in S$:*

$$V^* \succeq V_{\sigma,\gamma}. \tag{230}$$

Proof

We know that for all $w \in S$:

$$V_{\sigma,\gamma}(w) = R(w) + \gamma \mathbf{E}_{w' \in T_\sigma(w)}[V_{\sigma,\gamma}(w')] \tag{231}$$

Applying $\mathbf{V}_{\sup}$ yields:

$$\mathbf{V}_{\sup}(V_{\sigma,\gamma})(w) = R(w) + \gamma \sup_{a \in A(w)} \mathbf{E}_{w' \in a}[R(w')] \tag{232}$$

Because $T_\sigma(w)$ is a particular $a \in A(w)$:

$$\mathbf{V}_{\sup}(V_{\sigma,\gamma})(w) \geq R(w) + \gamma \mathbf{E}_{w' \in T_\sigma(w)}[V_{\sigma,\gamma}(w')] \tag{233}$$

$$\mathbf{V}_{\sup}(V_{\sigma,\gamma})(w) \geq V_{\sigma,\gamma}(w) \tag{234}$$

Thus, $\mathbf{V}_{\sup}(V_{\sigma,\gamma}) \succeq V_{\sigma,\gamma}$. By Lemma 60, $V^* \succeq V_{\sigma,\gamma}$. ■

Theorem 62 *If V_γ^* is the fixed point of $\mathbf{V}_{\sup}$ for γ, R is Lipschitz continuous, then for any σ where T_σ is a contraction mapping, if $\lim_{\gamma \to 1^-}(1-\gamma)V_\gamma^*$ exists, then*

$$\lim_{\gamma \to 1^-}(1-\gamma)V_\gamma^* \succeq \bar{V}_\sigma. \tag{235}$$

Proof By Theorem 54, for all w, $\lim_{\gamma \to 1^-}(1-\gamma)V_{\sigma,\gamma}(w) = \bar{V}_\sigma(w)$. By Theorem 61, $V_\gamma^* \succeq V_{\sigma,\gamma}$. Finally, we use the fact that if, for all x, $f(x) \geq g(x)$, then $\lim_{x \to c^-} f(x) \geq \lim_{x \to c^-} g(x)$. ■

Theorem 63 *If V_γ^* is the fixed point of $\mathbf{V}_{\sup}$ for γ, R is Lipschitz continuous, if $\lim_{\gamma \to 1^-}(1-\gamma)V_\gamma^*$ exists, then for any σ where T_σ is a contraction mapping, if $f : P(S,d) \to P(M,d)$ is an extension of T_σ which is a contraction mapping, then there exists a $D^* \in P(S,d)$ where $f(D^*) = D^*$, and:*

$$\lim_{\gamma \to 1^-}(1-\gamma)V_\gamma^*(w) \geq \mathbf{E}_{w \in D^*}[R(w)] \tag{236}$$

Proof By Theorem 62:

$$\lim_{\gamma \to 1^-}(1-\gamma)V_\gamma^* \succeq \bar{V}_\sigma. \tag{237}$$

Also by Theorem 53, $\bar{V}_\sigma = \lim_{t \to \infty} r_t(\sigma, w)$. By definition, $\lim_{t \to \infty} \mathbf{E}_{w \in T_\sigma^t}[R(w)]$. By Theorem 35, $\lim_{t \to \infty} \mathbf{E}_{w \in T_\sigma^t}[R(w)] = \mathbf{E}_{w \in D^*}[R(w)]$. The result follows by combining these bounds.

■

I Limiting the Squared Difference From Optimal

We want to bound the expected squared distance of the stationary distribution D_η^* from the optimal point. Without loss of generality, assume $w^* = 0$. If we define $R(w) = w^2$, then $\mathbf{E}_{w \in D_\eta^*}[R(w)]$ is the value we want to bound. Next, we define $A(w)$ such that $\mathbf{p}(w) \in A(w)$.

Instead of tying the proof too tightly to gradient descent, we consider arbitrary real-valued parameters M, K, $r \in [0, 1)$. We define $S = \{w \in \mathbf{R}^n : \|w\| \leq K\}$. For all w, define $A(w)$ to be the set of all discrete distributions $X \in D(S, d)$ such that:

1. $E[X \cdot w] \leq (1-r)w \cdot w$, and
2. $\|X - (1-r)w\| \leq M$.

We wish to calculate the maximum expected squared value of this process. In particular, this can be represented as an infinite horizon average reward MDP, where the reward at a state is w^2. We know that zero is a state reached in the optimal solution. Thus, we are concerned with bounding $V^*(0)$.

Define $A(w)$ to be the set of random variables such that for all random variables $a \in A(w)$:

$$|a| \leq M \tag{238}$$
$$\mathbf{E}_{x \in a}[x \cdot w] \leq 0 \tag{239}$$

The Bellman equation, given a discount factor γ, is:

$$V^*_\gamma(w) = w^2 + \gamma \sup_{a \in A(w)} \mathbf{E}[V^*_\gamma(a)] \tag{240}$$

We can relate this bound on the value to any stationary distribution.

Theorem 64 *If $\mathbf{p} : P(S,d) \to P(S,d)$ is a contraction mapping such that for all $w \in S$, $\mathbf{p}(I_w) \in A(w)$,*

then there exists a unique $D^* \in P(S,d)$ where $\mathbf{p}(D^*) = D^*$, and:

$$\lim_{\gamma \to 1^-} (1-\gamma) V^*_\gamma(w) \geq \mathbf{E}_{w \in D^*}[w^2] \tag{241}$$

This follows directly from Theorem 63.

Theorem 65 *The solution to the Bellman equation (Equation 240) is:*

$$V^*_\gamma(w) = \frac{1}{1-\gamma(1-r)^2}\left(w^2 + \frac{\gamma}{1-\gamma}M^2\right) \tag{242}$$

Proof In order to distinguish between the question and the answer, we write the candidate from Equation 242:

$$V_\gamma = \frac{1}{1-\gamma(1-r)^2}\left(w^2 + \frac{\gamma}{1-\gamma}M^2\right) \tag{243}$$

Therefore, we are interested in discovering what the Bellman operator does to V_γ. First of all, define $B(w)$ to be the set of random variables such that for all random variables $b \in B(w)$:

$$|b| \leq M \tag{244}$$
$$\mathbf{E}_{x \in b}[x \cdot w] \leq 0 \tag{245}$$

Thus, for every $a \in A(w)$, there exists a $b \in B(w)$ such that $a = (1-r)w + b$, and for every $b \in B(w)$, there exists an $a \in A(w)$ such that $a = (1-r)w + b$. Therefore,

$$\sup_{a \in A(w)} \mathbf{E}[V_\gamma(a)] = \sup_{a \in B(w)} \mathbf{E}[V_\gamma((1-r)w + a)] \tag{246}$$
$$= \frac{1}{1-\gamma(1-r)^2}\frac{\gamma}{1-\gamma}M^2 + \frac{1}{1-\gamma(1-r)^2}\sup_{a \in B(w)} \mathbf{E}[((1-r)w + a)^2] \tag{247}$$

Expanding the last part:

$$\sup_{a \in B(w)} \mathbf{E}[((1-r)w + a)^2] = \sup_{a \in B(w)} (1-r)^2 w^2 + 2(1-r)\mathbf{E}[w \cdot a] + \mathbf{E}[a^2] \tag{248}$$

By Equation (238):

$$\sup_{a \in B(w)} \mathbf{E}[((1-r)w + a)^2] \leq \sup_{a \in B(w)} (1-r)^2 w^2 + 2(1-r)\mathbf{E}[w \cdot a] + M^2 \tag{249}$$

By Equation (239):

$$\sup_{a \in B(w)} \mathbf{E}[((1-r)w + a)^2] \leq \sup_{a \in B(w)} (1-r)^2 w^2 + M^2 \tag{250}$$

$$\sup_{a \in B(w)} \mathbf{E}[((1-r)w + a)^2] \leq (1-r)^2 w^2 + M^2 \tag{251}$$

Also, note that if $\Pr[a = \frac{M}{\|w\|} w] = \Pr[a = -\frac{M}{\|w\|} w] = 0.5$, then

$$\mathbf{E}[((1-r)w + a)^2] = ((1-r)w + M)^2 + ((1-r)w - M)^2 \tag{252}$$

$$= (1-r)^2 w^2 + M^2. \tag{253}$$

Thus, $\sup_{a \in A(w)} \mathbf{E}[((1-r)w + a)^2] = (1-r)^2 w^2 + M^2$. Plugging this into Equation (247):

$$\sup_{a \in A(w)} \mathbf{E}[V_\gamma(a)] = \frac{1}{1-\gamma(1-r)^2} \frac{\gamma}{1-\gamma} M^2 + \frac{1}{1-\gamma(1-r)^2} \left((1-r)^2 w^2 + M^2\right) \tag{254}$$

$$= \frac{1}{1-\gamma(1-r)^2} \frac{1}{1-\gamma} M^2 + \frac{1}{1-\gamma(1-r)^2} (1-r)^2 w^2 \tag{255}$$

Plugging this into the recursion yields:

$$w^2 + \gamma \sup_{a \in A(w)} \mathbf{E}[V_\gamma(a)] = w^2 + \gamma \left(\frac{1}{1-\gamma(1-r)^2} \frac{1}{1-\gamma} M^2 + \frac{1}{1-\gamma(1-r)^2} (1-r)^2 w^2 \right) \tag{256}$$

$$w^2 + \gamma \sup_{a \in A(w)} \mathbf{E}[V_\gamma(a)] = \frac{1}{1-\gamma(1-r)^2} w^2 + \frac{1}{1-\gamma(1-r)^2} \frac{\gamma}{1-\gamma} M^2 \tag{257}$$

$$w^2 + \gamma \sup_{a \in A(w)} \mathbf{E}[V_\gamma(a)] = V_\gamma(w) \tag{258}$$

Therefore, V_γ satisfies the supremum Bellman equation. ■

Theorem 66 *If $\mathbf{p} : P(S,d) \to P(S,d)$ is a contraction mapping such that for all $w \in S$, $\mathbf{p}(I_w) \in A(w)$,*

then there exists a unique $D^* \in P(S,d)$ where $\mathbf{p}(D^*) = D^*$, and:

$$E_{w \in D^*}[w^2] \leq \frac{M^2}{(2-r)r} \tag{259}$$

Proof By Theorem 64:

$$E_{w \in D^*}[w^2] \leq \lim_{\gamma \to 1^-} (1-\gamma) V_\gamma^*(w) \tag{260}$$

By Theorem 65, for any w:

$$E_{w \in D^*}[w^2] \leq \lim_{\gamma \to 1^-} (1-\gamma) \frac{1}{1-\gamma(1-r)^2} \left(w^2 + \frac{\gamma}{1-\gamma} M^2 \right) \tag{261}$$

$$E_{w \in D^*}[w^2] \leq \lim_{\gamma \to 1^-} \frac{1}{1-\gamma(1-r)^2} \left((1-\gamma) w^2 + \gamma M^2\right) \tag{262}$$

$$E_{w \in D^*}[w^2] \leq \frac{1}{1-(1)(1-r)^2} \left(0(w^2) + 1(M^2)\right) \tag{263}$$

$$E_{w \in D^*}[w^2] \leq \frac{M^2}{1-(1-r)^2} \tag{264}$$

$$E_{w \in D^*}[w^2] \leq \frac{M^2}{(2-r)r} \tag{265}$$

■

Theorem 10 *The average squared distance of the stationary distance from the optimal point is bounded by:*

$$\frac{4\eta G^2}{(2-\eta\lambda)\lambda}.$$

In other words, the squared distance is bounded by $O(\eta G^2/\lambda)$.

Proof

By Theorem 42 and Theorem 43, the stationary distribution of the stochastic process satisfies the constraints of Theorem 66 with $r = \eta\lambda$ and $M = 2\eta G$. Thus, substituting into Theorem 66 yields the result. ■

J Application to Stochastic Gradient Descent

An SVM has a cost function consisting of regularization and loss:

$$c(w) = \frac{\lambda}{2}w^2 + \frac{1}{m}\sum_{i=1}^{m} L(y^i, w^i \cdot x^i) \tag{266}$$

In this section, we assume that we are trying to find the optimal weight vector given an SVM:

$$\operatorname*{argmin}_w c(w) \tag{267}$$

In the following, we assume $y^i \in \{-1, +1\}$, $x^i \cdot x^i = 1$, and $L(y, \hat{y}) = \frac{1}{2}(\max(1 - y\hat{y}, 0))^2$ is convex in $\hat{y}$, and $\frac{\partial L(y,\hat{y})}{\partial \hat{y}}$ is Lipschitz continuous. At each time step, we select an i uniformly at random between 1 and m and take a gradient step with respect to:

$$c^i(w) = \frac{\lambda}{2}w^2 + L(y^i, w \cdot x^i) \tag{268}$$

Define $f^i(w) = L(y^i, w \cdot x^i)$. In other words:

$$\nabla c^i(w) = \lambda w + \nabla f^i(w) \tag{269}$$

This results in the update:

$$w^{t+1} = w^t - \eta(\lambda w^t + \nabla f^i(w)) \tag{270}$$

In our case, $\nabla f^i(w) = x^i \frac{\partial}{\partial \hat{y}} L(y^i, \hat{y})$. Define ϕ^i such that:

$$\phi^i(w) = w - \eta(\lambda w + \nabla f^i(w)) \tag{271}$$

In what will follow, we assume that $\left\|\nabla f^i(w)\right\|$ and $\left\|\nabla f^i(w)\right\|_{\text{Lip}}$ are both bounded. This will require bounds on $\left\|x^i\right\|$.

In the first section, we analyze how stochastic gradient descent is a contraction mapping. In the second section, we analyze the implications of this result.

K Putting it all Together

Theorem 67

$$\sigma_{D_\eta^*} \leq \frac{2\sqrt{\eta}G}{\sqrt{(2-\eta\lambda)\lambda}} \tag{272}$$

Corollary 68 *If $\eta \leq \eta^*$, then $(1 - \eta\lambda) \geq 0$, and:*

$$\sigma_{D_\eta^*} \leq \frac{2\sqrt{\eta}G}{\sqrt{\lambda}} \tag{273}$$

Proof By Theorem 31, $\sigma_{D_\eta^*}^{w^*} \geq \sigma_{D_\eta^*}$. The result follows from Theorem 10. ■

Define D_η^t to be the distribution of the stochastic gradient descent update after t iterations, and D_η^0 to be the initial distribution.

Theorem 69 *If $w_0 = 0$, then $W_2(D_\eta^0, D_\eta^*) \leq \frac{G}{\lambda}$, and $W_1(D_\eta^0, D_\eta^*) \leq \frac{G}{\lambda}$.*

Proof By Theorem 44, $\|w_t\| \leq \frac{G}{\lambda}$. Therefore, for any $w \in D_\eta^*$, $\|w\| \leq \frac{G}{\lambda}$. The result follows directly. ■

Theorem 70 *If D_η^t is the distribution of the stochastic gradient descent update after t iterations, and $\eta \leq \eta^*$, then:*

$$d(\mu_{D_\eta^t}, \mu_{D_\eta^*}) \leq \frac{G}{\lambda}(1 - \eta\lambda)^t \tag{274}$$

$$\sigma_{D_\eta^t} \leq \sigma_{D_\eta^*} + \frac{G}{\lambda}(1 - \eta\lambda)^t \tag{275}$$

Corollary 71 *If $w_0 = 0$, then by Theorem 69 and Corollary 68:*

$$d(\mu_{D_\eta^t}, \mu_{D_\eta^*}) \leq \frac{G}{\lambda}(1 - \eta\lambda)^t \tag{276}$$

$$\sigma_{D_\eta^t} \leq \frac{2\sqrt{\eta}G}{\sqrt{\lambda}} + \frac{G}{\lambda}(1 - \eta\lambda)^t \tag{277}$$

Proof

Note that by Theorem 8:

$$W_1(D_\eta^t, D_\eta^*) \leq \frac{G}{\lambda}(1 - \eta\lambda)^t. \tag{278}$$

Equation 274 follows from Corollary 23.

Similarly by Theorem 8:

$$W_2(D_\eta^t, D_\eta^*) \leq W_2(D_\eta^0, D_\eta^*)(1 - \eta\lambda)^t. \tag{279}$$

Equation 275 follows from Theorem 32. ■

Theorem 11 *Given a cost function c such that $\|c\|_{\text{Lip}}$ and $\|\nabla c\|_{\text{Lip}}$ are bounded, a distribution D such that σ_D and is bounded, then, for any v:*

$$\begin{aligned}\mathbf{E}_{w \in D}[c(w)] - \min_w c(w) \leq & (\sigma_D^v)\sqrt{2\|\nabla c\|_{\text{Lip}}(c(v) - \min_w c(w))} \\ & + \frac{\|\nabla c\|_{\text{Lip}}}{2}(\sigma_D^v)^2 + (c(v) - \min_w c(w))\end{aligned} \tag{280}$$

Proof First, we observe that, for any w', since ∇c is Lipschitz continuous:

$$c(w') - c(v) = \int_{a \in [0,1]} \nabla c(a(w' - v)) + v) \cdot (w' - v) da \tag{281}$$

For any w'', by definition of Lipschitz continuity $\|\nabla c(w'') - \nabla c(v)\| \leq \|\nabla c\|_{\text{Lip}} \|w'' - v\|$, so by the triangle inequality:

$$\|\nabla c(w'')\| - \|\nabla c(v)\| \leq \|\nabla c\|_{\text{Lip}} \|w'' - v\| \tag{282}$$

Applying this to $a(w' - v) + v$ for $a \in [0, 1]$ yields:

$$\|\nabla c(a(w' - v) + v)\| - \|\nabla c(v)\| \leq \|\nabla c\|_{\text{Lip}} \|a(w' - v)\| \tag{283}$$

$$\|\nabla c(a(w' - v) + v)\| - \|\nabla c(v)\| \leq \|\nabla c\|_{\text{Lip}} a \|(w' - v)\| \tag{284}$$

Thus, by the Cauchy-Schwartz inequality:

$$\nabla c(a(w' - v) + v) \cdot (w' - v) \leq (\|\nabla c\|_{\text{Lip}} a \|w' - v\| + \|\nabla c(v)\|) \|w' - v\|. \tag{285}$$

If f, g are integrable, real valued functions, and if $f(x) \leq g(x)$ for all $x \in [a, b]$, then $\int_a^b f(x)dx \leq \int_a^b g(x)dx$. Therefore:

$$c(w') - c(v) \leq \int_{a \in [0,1]} (\|\nabla c\|_{\text{Lip}} a \|w' - v\| + \|\nabla c(v)\|) \|w' - v\| \, da \tag{286}$$

$$c(w') - c(v) \leq (\frac{1}{2} \|\nabla c\|_{\text{Lip}} \|w' - v\| + \|\nabla c(v)\|) \|w' - v\| \tag{287}$$

$$c(w') - c(v) \leq \frac{1}{2} \|\nabla c\|_{\text{Lip}} (\|w' - v\|)^2 + \|\nabla c(v)\|) \|w' - v\| \tag{288}$$

We break this down into three pieces: $c_2(w') = \frac{1}{2} \|\nabla c\|_{\text{Lip}} (\|w' - v\|)^2$, $c_1(w') = \|\nabla c(v)\| \|w' - v\|$, and $c_0(w') = c(v)$ (i.e.c_0 is constant). Therefore:

$$c(w') \leq c_0(w') + c_1(w') + c_2(w') \tag{289}$$

By Corollary 25 and $\|c_1\|_{\text{Lip}} = \|\nabla c(v)\|$:

$$\mathbf{E}_{w' \in D}[c_1(w')] - c_1(v) \leq \|c_1\|_{\text{Lip}} W_1(D, v) \tag{290}$$

Note that $\|c_2\|_{L_2} = \frac{1}{2} \|\nabla c\|_{\text{Lip}}$ Using the extension of Kantorovich-Rubinstein:

$$\mathbf{E}_{w' \in D}[c_2(w')] - c_2(v) \leq \|c_2\|_{L_2} (W_2(D, v))^2 \tag{291}$$

Because c_0 is a constant function:

$$\mathbf{E}_{w' \in D}[c_0(w')] - c_0(v) = 0 \tag{292}$$

Thus, putting it together:

$$\mathbf{E}_{w' \in D}[c(w')] - c(v) \leq \|c_2\|_{L_2} (W_2(D, v))^2 + \|c_1\|_{\text{Lip}} W_1(D, v) \tag{293}$$

$$\mathbf{E}_{w' \in D}[c(w')] - c(v) \leq \frac{1}{2} \|\nabla c\|_{\text{Lip}} (W_2(D, v))^2 + \|\nabla c(v)\| W_1(D, v) \tag{294}$$

Since by Fact 28, $W_2(D, v) = \sigma_D^v$, and by Theorem 33, $W_2(D, v) \geq W_1(D, v)$, so:

$$\mathbf{E}_{w' \in D}[c(w')] - c(v) \leq \frac{1}{2} \|\nabla c\|_{\text{Lip}} (\sigma_D^v)^2 + \|\nabla c(v)\| \sigma_D^v \tag{295}$$

By Theorem 13:

$$\|\nabla c(v)\| \leq \sqrt{2 \|\nabla c\|_{\text{Lip}} [c(v) - \min_w c(w)]}. \tag{296}$$

$$\mathbf{E}_{w' \in D}[c(w')] - c(v) \leq \frac{1}{2} \|\nabla c\|_{\text{Lip}} (\sigma_D^v)^2 + \sigma_D^v \sqrt{2 \|\nabla c\|_{\text{Lip}} [c(v) - \min_w c(w)]} \tag{297}$$

Adding $c(v) - \min_w c(w)$ to both sides yields the result. ■

Theorem 72 *If $\eta \leq \eta^*$ and $T = \frac{\ln k - (\ln \eta + \ln \lambda)}{2\eta\lambda}$:*

$$\mathbf{E}_{w \in D_\eta^{T,k}}[c(w)] - \min_w c(w) \leq \frac{8\eta G^2}{\sqrt{k\lambda}} \sqrt{\|\nabla c\|_{\text{Lip}}} + \frac{8\eta G^2 \|\nabla c\|_{\text{Lip}}}{k\lambda} + (2\eta G^2). \tag{298}$$

Proof Define $D_\eta^{T,k}$ to be the average of k draws from D_η^T. By Theorem 34:

$$\mu_{D_\eta^{T,k}} = \mu_{D_\eta^T} \tag{299}$$

$$\sigma_{D_\eta^{T,k}} = \frac{1}{\sqrt{k}} \sigma_{D_\eta^T} \tag{300}$$

Applying Corollary 71:

$$d(\mu_{D_\eta^{T,k}}, \mu_{D_\eta^*}) \leq \frac{G}{\lambda}(1 - \eta\lambda)^T \tag{301}$$

$$\sigma_{D_\eta^{T,k}} \leq \frac{1}{\sqrt{k}} \left(\frac{2\sqrt{\eta} G}{\sqrt{\lambda}} + \frac{G}{\lambda}(1 - \eta\lambda)^T \right) \tag{302}$$

Since $1 - \eta\lambda \in [0, 1]$, $\exp(-\eta\lambda) \leq 1 - \eta\lambda$, so:

$$d(\mu_{D_\eta^{T,k}}, \mu_{D_\eta^*}) \leq \frac{G}{\lambda} \exp(-\eta\lambda T) \tag{303}$$

$$\sigma_{D_\eta^{T,k}} \leq \frac{1}{\sqrt{k}} \left(\frac{2\sqrt{\eta} G}{\sqrt{\lambda}} + \frac{G}{\lambda} \exp(-\eta\lambda T) \right) \tag{304}$$

Note that $\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}} \leq \sigma_{D_\eta^{T,k}} + d(\mu_{D_\eta^{T,k}}, \mu_{D_\eta^*})$. So:

$$\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}} \leq \frac{1}{\sqrt{k}} \left(\frac{2\sqrt{\eta} G}{\sqrt{\lambda}} + \frac{G}{\lambda} \exp(-\eta\lambda T) \right) + \frac{G}{\lambda} \exp(-\eta\lambda T) \tag{305}$$

$$\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}} \leq \frac{2\sqrt{\eta} G}{\sqrt{k\lambda}} + \frac{2G}{\lambda} \exp(-\eta\lambda T) \tag{306}$$

Setting $T = \frac{\ln k - (\ln \eta + \ln \lambda)}{2\eta\lambda}$ yields:

$$\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}} \leq \frac{4\sqrt{\eta} G}{\sqrt{k\lambda}} \tag{307}$$

By Theorem 11:

$$\mathbf{E}_{w \in D_\eta^{T,k}}[c(w)] - \min_w c(w) \leq (\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}}) \sqrt{2 \|\nabla c\|_{\text{Lip}} (c(\mu_{D_\eta^*}) - \min_w c(w))} + \frac{\|\nabla c\|_{\text{Lip}}}{2} (\sigma_{D_\eta^{T,k}}^{\mu_{D_\eta^*}})^2 + (c(\mu_{D_\eta^*}) - \min_w c(w)) \tag{308}$$

$$\leq \frac{4\sqrt{\eta} G}{\sqrt{k\lambda}} \sqrt{2 \|\nabla c\|_{\text{Lip}} (c(\mu_{D_\eta^*}) - \min_w c(w))} + \frac{\|\nabla c\|_{\text{Lip}}}{2} \frac{16\eta G^2}{k\lambda} + (c(\mu_{D_\eta^*}) - \min_w c(w)). \tag{309}$$

By Theorem 9, $c(\mu_{D_\eta^*}) - \min_w c(w) \leq 2\eta G^2$:

$$\mathbf{E}_{w \in D_\eta^{T,k}}[c(w)] - \min_w c(w) \leq \frac{4\sqrt{\eta} G}{\sqrt{k\lambda}} \sqrt{2 \|\nabla c\|_{\text{Lip}} (2\eta G^2)} + \frac{\|\nabla c\|_{\text{Lip}}}{2} \frac{16\eta G^2}{k\lambda} + (2\eta G^2) \tag{310}$$

$$\leq \frac{8\eta G^2}{\sqrt{k\lambda}} \sqrt{\|\nabla c\|_{\text{Lip}}} + \frac{8\eta G^2 \|\nabla c\|_{\text{Lip}}}{k\lambda} + (2\eta G^2). \tag{311}$$