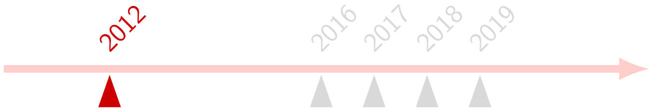
# Online Step Size Adaptation
## for Stochastic Optimization

Andrii Zadaiancuk
Universität Tübingen
January 27, 2019

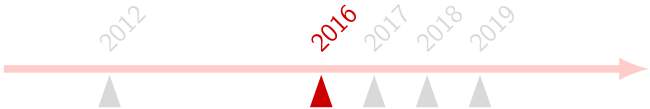
# In My Timeline




- October 2012: Bachalor in Applied Mathematics and Physics, Moscow, Russia
- June 2016: **Probabilistic Pruning of Neural Networks**, Bachelor Thesis and Publication under supervision Prof. Dr. Vadim Strijov
- October 2016: Neural Information Processing, Tübingen, Germany
- September 2017 - November 2017: Architectures of Generative Nets, Bethge Lab
- November 2017 - January 2018: Probabilistic RPROP, Probabilistic Numerics, MPI IS
- February 2018 - August 2018: **Proximal step-size adaptation**, Master Thesis under Prof. Dr. Phillipp Henig supervision
- October 2018 - March 2019: Equation Learning for extrapolation in Model-based RL, Autonomus Learning Group, MPI IS

# In My Timeline

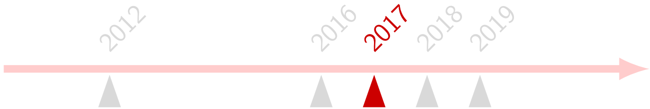





- October 2012: Bachalor in Applied Mathematics and Physics, Moscow, Russia
- June 2016: **Probabilistic Pruning of Neural Networks**, Bachelor Thesis and Publication under supervision Prof. Dr. Vadim Strijov
- October 2016: Neural Information Processing, Tübingen, Germany
- September 2017 - November 2017: Architectures of Generative Nets, Bethge Lab
- November 2017 - January 2018: Probabilistic RPROP, Probabilistic Numerics, MPI IS
- February 2018 - August 2018: **Proximal step-size adaptation**, Master Thesis under Prof. Dr. Phillipp Henig supervision
- October 2018 - March 2019: Equation Learning for extrapolation in Model-based RL, Autonomus Learning Group, MPI IS

# In My Timeline

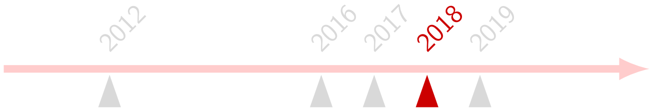





- October 2012: Bachalor in Applied Mathematics and Physics, Moscow, Russia
- June 2016: **Probabilistic Pruning of Neural Networks**, Bachelor Thesis and Publication under supervision Prof. Dr. Vadim Strijov
- October 2016: Neural Information Processing, Tübingen, Germany
- September 2017 - November 2017: Architectures of Generative Nets, Bethge Lab
- November 2017 - January 2018: Probabilistic RPROP, Probabilistic Numerics, MPI IS
- February 2018 - August 2018: **Proximal step-size adaptation**, Master Thesis under Prof. Dr. Phillipp Henig supervision
- October 2018 - March 2019: Equation Learning for extrapolation in Model-based RL, Autonomus Learning Group, MPI IS

# In My Timeline

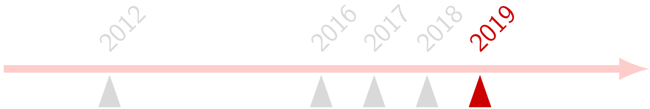





- October 2012: Bachalor in Applied Mathematics and Physics, Moscow, Russia
- June 2016: **Probabilistic Pruning of Neural Networks**, Bachelor Thesis and Publication under supervision Prof. Dr. Vadim Strijov
- October 2016: Neural Information Processing, Tübingen, Germany
- September 2017 - November 2017: Architectures of Generative Nets, Bethge Lab
- November 2017 - January 2018: Probabilistic RPROP, Probabilistic Numerics, MPI IS
- February 2018 - August 2018: **Proximal step-size adaptation**, Master Thesis under Prof. Dr. Phillipp Henig supervision
- October 2018 - March 2019: Equation Learning for extrapolation in Model-based RL, Autonomus Learning Group, MPI IS

# In My Timeline

EBERHARD KARLS UNIVERSITÄT TÜBINGEN




impr-is

- October 2012: Bachalor in Applied Mathematics and Physics, Moscow, Russia
- June 2016: **Probabilistic Pruning of Neural Networks**, Bachelor Thesis and Publication under supervision Prof. Dr. Vadim Strijov
- October 2016: Neural Information Processing, Tübingen, Germany
- September 2017 - November 2017: Architectures of Generative Nets, Bethge Lab
- November 2017 - January 2018: Probabilistic RPROP, Probabilistic Numerics, MPI IS
- February 2018 - August 2018: **Proximal step-size adaptation**, Master Thesis under Prof. Dr. Phillipp Henig supervision
- October 2018 - March 2019: Equation Learning for extrapolation in Model-based RL, Autonomus Learning Group, MPI IS

# Outline


Problem formulation

Hypergradient Descent (HD) Adaptation
- Proximal Point Interpretation

Proximal Quadratic (PQ) Adaptation
- Bias of the Minimum of Quadratic Model
- Proximal Point Iteration for Quadratic Model

Experiments
- Fine-tunned adaptation models
- Sensitivity to the hyperparameters

Conclusions

# Stochastic Optimization in Machine Leaning

## Regularized empirical risk minimization

$$\min_{\theta} R_{emp}(\theta) + \mathcal{L}_{reg}$$

$$R_{emp}(\theta) = \frac{1}{N} \sum_{i=1}^{N} l(h(x_i, \theta), y_i)$$

## Stochastic optimization

SGD update rule

$$\theta_{t+1}(\alpha) = \theta_t - \alpha g(\theta_t),$$

where $g(\theta)$ is the stochastic gradient

$$g(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla l(\theta; x_i).$$

# Online Step Size Adaptation

Parameter update is

$$\theta_{t+1} = \theta_t + \alpha v_t.$$

Optimal step size for iteration t is equal to

$$\alpha_t^* = \arg\min_{\alpha} \mathcal{L}(\theta_{t+1}(\alpha)).$$

In case when it is too expensive too find the exact minimum of the loss function (e.g. by line search), one can **adapt** the previous step size value $\alpha_{t-1}$ to make it closer to the $\alpha_t^*$.

# Hypergradient Descent (HD) Adaptation

$$\theta_{t+1} = \theta_t + \alpha_t v_t.$$

We want to make $\alpha_t$ step closer to the optimal $\alpha_t^*$

$$\alpha_t = \alpha_{t-1} - \beta \frac{\partial \mathcal{L}(\theta_{t+1})}{\partial \alpha}.$$

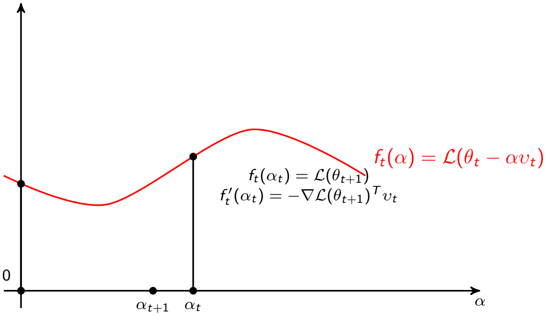
Using chain rule

$$\alpha_t = \alpha_{t-1} - \beta \nabla_\theta \mathcal{L}(\theta_{t+1})^T v_t.$$

As gradient $\nabla_\theta \mathcal{L}(\theta_{t+1})$ is unknown during step $t$, we assume that $\alpha_t^* \approx \alpha_{t-1}^*$

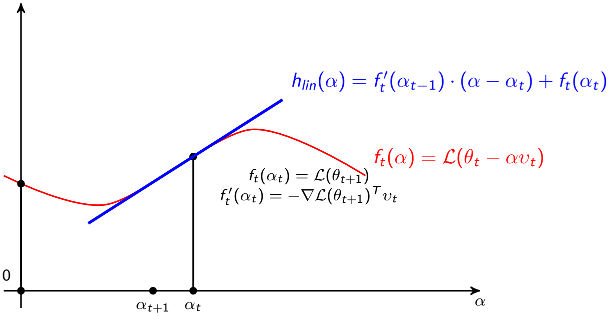
$$\alpha_t = \alpha_{t-1} - \beta \nabla_\theta \mathcal{L}(\theta_t)^T v_{t-1}.$$

# HD as Proximal Point Iteration of Linear Model




Hypergradient Descent adaptation as iteration of Proximal Point algorithm applied to the linear model. We locally approximate $f_t(\alpha)$ by the linear model $h_{lin}(\alpha)$.

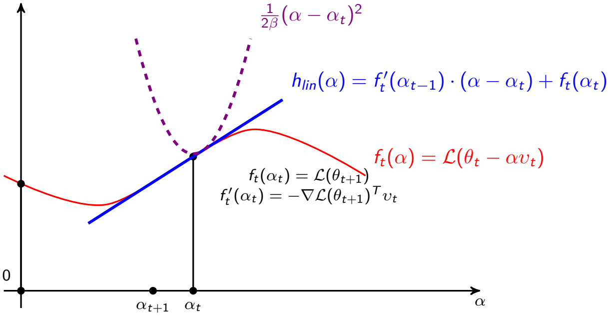
# HD as Proximal Point Iteration of Linear Model




Hypergradient Descent adaptation as iteration of Proximal Point algorithm applied to the linear model. We locally approximate $f_t(\alpha)$ by the linear model $h_{lin}(\alpha)$.

# HD as Proximal Point Iteration of Linear Model






Hypergradient Descent adaptation as iteration of Proximal Point algorithm applied to the linear model. We locally approximate $f_t(\alpha)$ by the linear model $h_{lin}(\alpha)$.

Proximal point iteration for model $h_t(\alpha)$:

$$\alpha_{t+1} = \arg\min_{\alpha} h_{lin}(\alpha) + \frac{1}{2\beta}(\alpha - \alpha_t)^2 = \alpha_t - \beta f_t'(\alpha_t).$$

# Proximal Adaptation



- Let us change the linear approximation $h_{lin}(\alpha)$ to another convex approximation $h(\alpha)$.
- As $h(\alpha)$ is **convex** and **one-dimensional**, we can easily compute its proximal operator.
- We can use one iteration of proximal point method for $h(\alpha)$ to adapt current step size $\alpha_t$.

Proximal step size adaptation with convex approximation $h(\alpha)$ of loss function $f(\alpha)$ is

$$\alpha_{t+1} = \arg\min_{\alpha} h(\alpha) + \frac{1}{2\beta}(\alpha - \alpha_t)^2.$$

# Proximal Quadratic (PQ) Adaptation

## Bias of the Minimum of Quadratic Model



Using Tailor expansion, we can get approximation of the expectation of the ratio of two random variables:

$$\mathbb{E}\left[\frac{X}{Y}\right] \approx \frac{\mu_x}{\mu_y} - \frac{Cov(X,Y)}{\mu_y^2} + \frac{\mathrm{var}[Y]\mu_x}{\mu_y^3}.$$

Applying it to our ratio we get

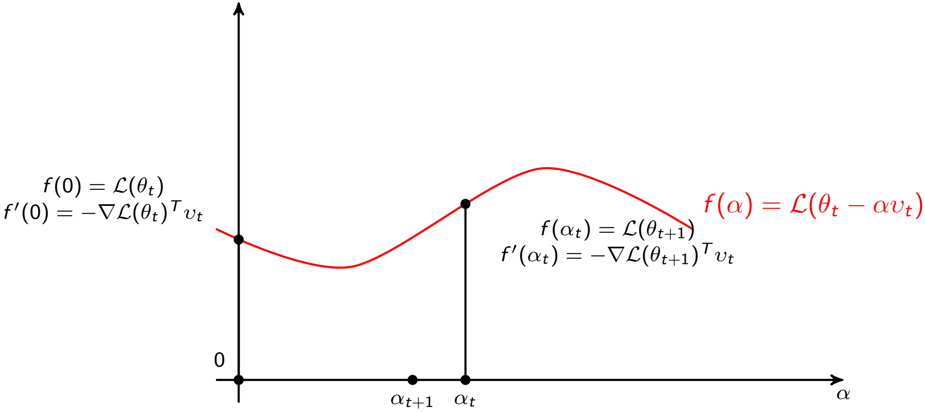
$$\mathbb{E}\left[\frac{\widehat{f}'(0)}{\widehat{f}'(\alpha_t) - \widehat{f}'(0)}\right] \approx \frac{f'(0)}{f'(\alpha_t) - f'(0)} + \underbrace{\frac{\sigma_0^2}{\left(f'(0) - f'(\alpha)\right)^2} + \frac{\left(\sigma_0^2 + \sigma_\alpha^2\right) f'(0)}{\left(f'(0) - f'(\alpha_t)\right)^3}}_{bias}.$$

We need to correct for this bias when the difference $f'(0) - f'(\alpha)$ is small or the noise in stochastic estimates $\widehat{f}'(\alpha)$ or $\widehat{f}'(0)$ is large.

# Proximal Quadratic (PQ) Adaptation

## Proximal Point Iteration for Quadratic Model

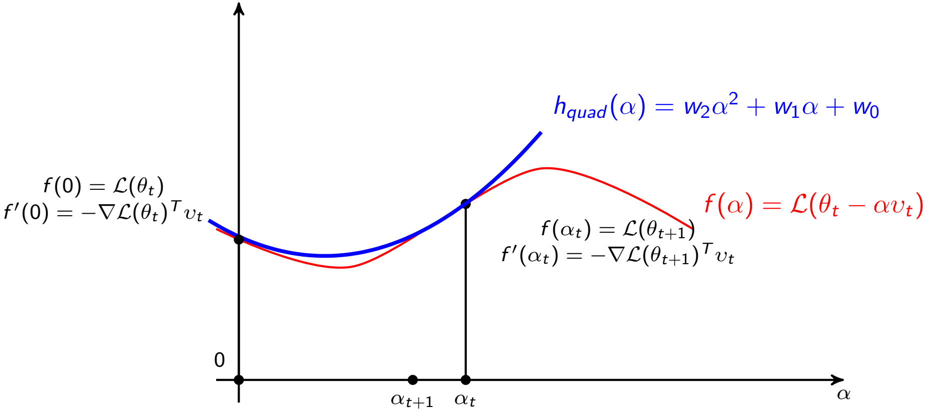





Proximal Quadratic (PQ) adaptation as iteration of Proximal Point algorithm applied to the quadratic model. We approximate $f_t(\alpha)$ by the quadratic model $h_{quad}(\alpha)$.

# Proximal Quadratic (PQ) Adaptation

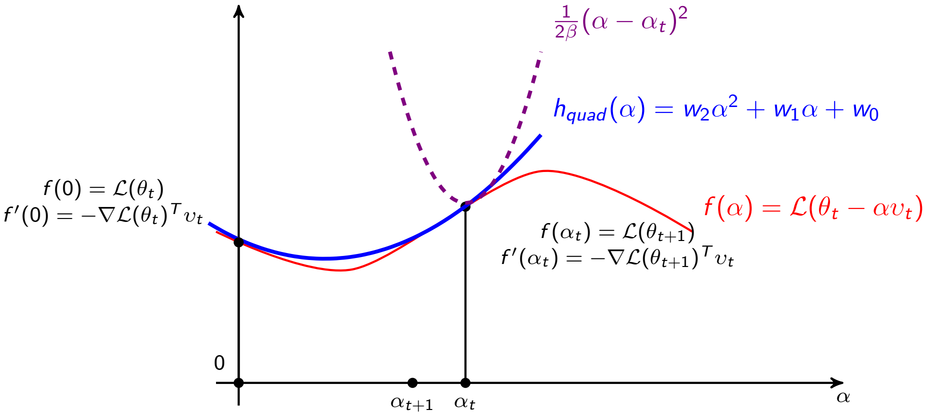
## Proximal Point Iteration for Quadratic Model




Proximal Quadratic (PQ) adaptation as iteration of Proximal Point algorithm applied to the quadratic model. We approximate $f_t(\alpha)$ by the quadratic model $h_{quad}(\alpha)$.

# Proximal Quadratic (PQ) Adaptation

## Proximal Point Iteration for Quadratic Model




Proximal Quadratic (PQ) adaptation as iteration of Proximal Point algorithm applied to the quadratic model. We approximate $f_t(\alpha)$ by the quadratic model $h_{quad}(\alpha)$.

# Proximal Quadratic (PQ) Adaptation

## Proximal Point Iteration for Quadratic Model

$$\operatorname{prox}_{\beta h_{quad}}(\alpha_t) = \arg\min_{\alpha} h_{quad}(\alpha) + \frac{1}{2\beta}(\alpha - \alpha_t)^2.$$

To find it we should take the derivative and set it to zero

$$\operatorname{prox}_{\beta h_{quad}}(\alpha_t) = \frac{\frac{1}{\beta}\alpha_t - w_1}{2w_2 + \frac{1}{\beta}}.$$

Step size update rule as one iteration of Proximal Point algorithm is

$$\alpha_{t+1} = \frac{\frac{1}{\beta}\alpha_t - w_1}{2w_2 + \frac{1}{\beta}}.$$

Using maximum-likelihood estimation $\widehat{w}$ of the parameters $w$

$$\widehat{\alpha}_{t+1} = \frac{\frac{1}{\beta}\alpha_t - \widehat{f}'(0)}{\frac{\widehat{f}'(\alpha) - \widehat{f}'(0)}{\alpha_t} + \frac{1}{\beta}}.$$

# PQ-Momentum Pseudocode



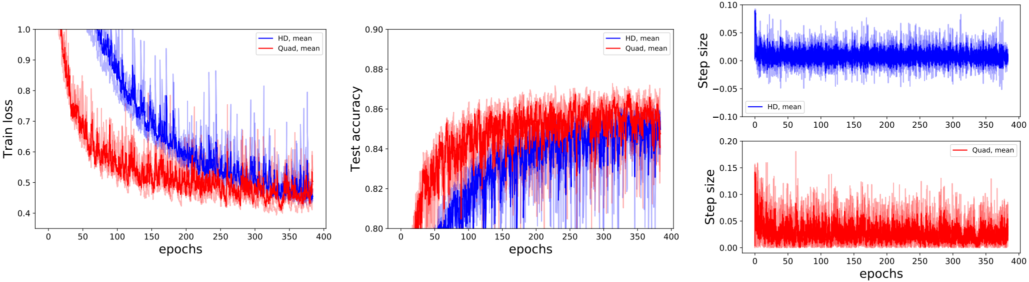
**Algorithm 1** Momentum with Proximal Quadratic adaptation (PQ-Momentum)

**Require:** initial parameter value $\theta_0$, initial step size $\alpha_0$, regularization constant $\beta$, momentum $\mu$, number of steps $T$, upper bound on Lipschitz constant $M$

Initialize $\upsilon = 0$, $m = 0$, $\alpha = \alpha_0$<br>
**for** $t = 1, \dots, T$ **do**<br>
  Evaluate stochastic gradient $g$<br>
  Evaluate one-dimentinal derivatives $\widehat{f}'(\alpha) = g^T \upsilon$ and $\widehat{f}'(0) = g_{old}^T \upsilon$<br>
  **if** $0 \leq \frac{\widehat{f}'(\alpha) - \widehat{f}'(0)}{\alpha} \leq M$ or $f'(0) > 0$ **then**<br>
    Update $\alpha = \frac{\frac{1}{\beta}\alpha - \widehat{f}'(0)}{\frac{\widehat{f}'(\alpha) - \widehat{f}'(0)}{\alpha} + \frac{1}{\beta}}$<br>
  **end if**<br>
  Update moving average $m = \mu m + (1 - \mu) g$<br>
  Evaluate new direction $\upsilon = -m$<br>
  Update parameters $\theta = \theta + \alpha \upsilon$<br>
  Update $g_{old} = g$<br>
**end for**

# Results

## Comparison of the fine-tunned adaptation models

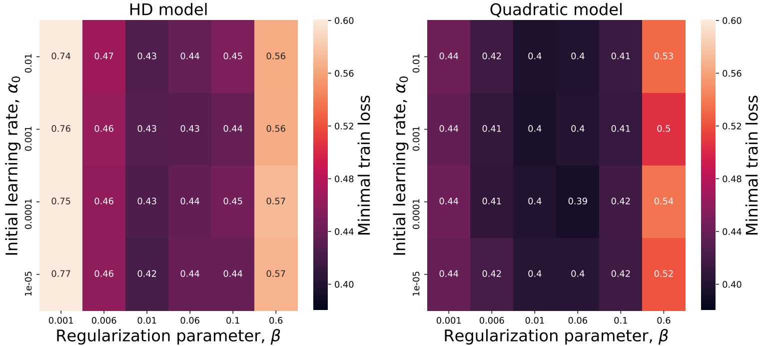



Experimental results of fine-tuned PQ and HD on CIFAR10. Parameter $\beta$ was chosen by grid search. PQ is superior to the HD. For both HD and PQ -Momentum the value of momentum parameter is $\mu = 0.99$.

# Results

## Sensitivity to the regularization parameter $\beta$






Sensitivity of the Proximal Quadratic and the Hypergradient Descent adaptation to initial step size $\alpha_0$ and regularization parameter $\beta$. Momentum with $\mu = 0.99$ on CIFAR10 with batch size 128.

# Conclusions

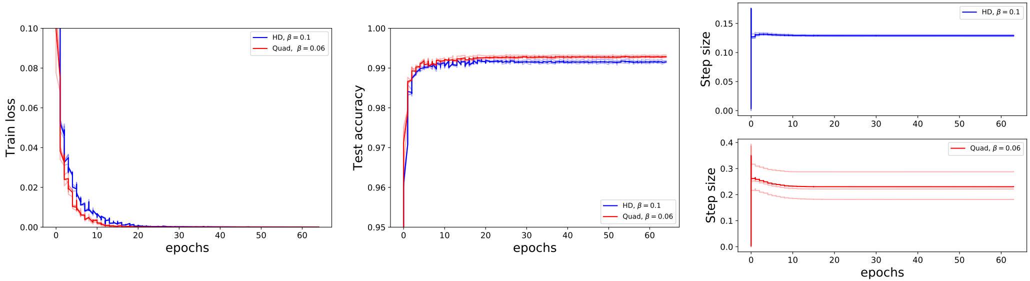


- Hypergradient Descent adaptation rule is equal to proximal point iteration of linear approximation.
- Proximal adaptation with **other convex approximation** is possible.
- Quadratic model **is biased** towards larger step sizes are therefore unstable.
- Proximal Quadratic adaptation **is less sensitive** to the hyperparameter choice.

Thank you for your attention!

# Results

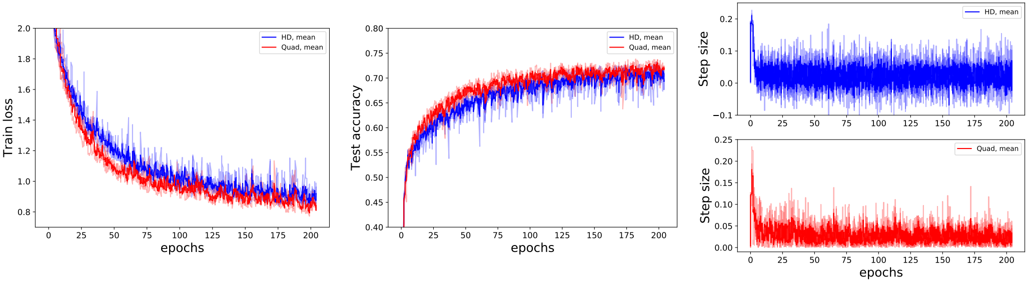
## Comparison of the fine-tunned adaptation models




Experimental results of fine-tuned PQ and HD on MNIST. Parameter $\beta$ was chosen by grid search. PQ is superior to the HD. For both HD and PQ -Momentum the value of momentum parameter is $\mu = 0.99$.

# Results

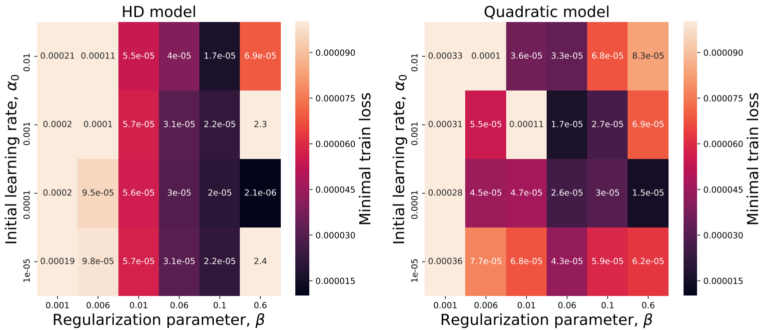
## Comparison of the fine-tuned adaptation models




Experimental results of fine-tuned PQ and HD on SVHN. Parameter $\beta$ was chosen by grid search. PQ is superior to the HD. For both HD and PQ -Momentum the value of momentum parameter is $\mu = 0.99$.

# Results

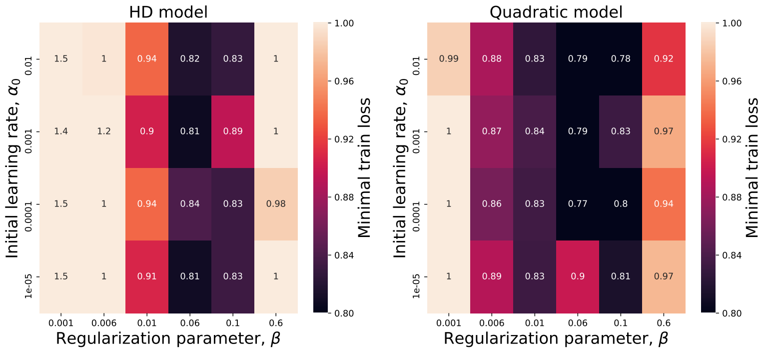
## Sensitivity to the regularization parameter $\beta$




Sensitivity of the Proximal Quadratic and the Hypergradient Descent adaptation to initial step size $\alpha_0$ and regularization parameter $\beta$. Momentum with $\mu = 0.99$ on MNIST with batch size 128.

# Results

## Sensitivity to the regularization parameter $\beta$




Sensitivity of the Proximal Quadratic and the Hypergradient Descent adaptation to initial step size $\alpha_0$ and regularization parameter $\beta$. Momentum with $\mu = 0.99$ on SVHN with batch size 128.